

# Multilayer Bootstrap Networks

Xiao-Lei Zhang

Received: date / Accepted: date

**Abstract** Multilayer bootstrap network builds a gradually narrowed multilayer nonlinear network from bottom up for unsupervised nonlinear dimensionality reduction. Each layer of the network is a group of  $k$ -centers clusterings. Each clustering uses randomly sampled data points with randomly selected features as its centers, and learns a one-of- $k$  encoding by one-nearest-neighbor optimization. Thanks to the binarized encoding, the similarity of two data points is measured by the number of the nearest centers they share in common, which is an adaptive similarity metric in the discrete space that needs no model assumption and parameter tuning. Thanks to the network structure, larger and larger local variations of data are gradually reduced from bottom up. The information loss caused by the binarized encoding is proportional to the correlation of the clusterings, both of which are reduced by the randomization steps.

**Keywords** Resampling · model ensembling · nearest neighbor · hierarchical model

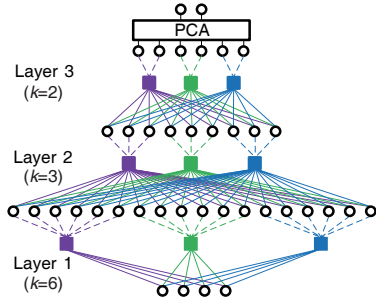
## 1 Introduction

Principle component analysis (PCA) is a simple and widely used unsupervised dimensionality reduction method, which finds a coordinate system that the linearly uncorrelated coordinate variables (called principle components) describe the most variances of data. Because PCA is insufficient to capture highly-nonlinear data distributions, nonlinear dimensionality reduction methods are explored. A nonlinear method faces two problems: the similarity between data points, and the definition of a nonlinear coordinate system.

Much effort has been made to define or learn a similarity measurement in a continuous space. Predefined similarity metrics include Euclidean distance and Gaussian radial basis functions (RBF) with tunable parameters. Learnable similarity metrics include Gaussian mixture model (GMM) with an expectation-maximization optimization. The former may not fit the true density of data, while the latter suffers from predefined model assumptions.

---

Xiao-Lei Zhang  
Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA  
E-mail: xiaolei.zhang9@gmail.com



**Fig. 1** Network structure. Each square represents a  $k$ -centers clustering.

Much work has been conducted on learning a nonlinear coordinate system. Examples include neural networks which build a piecewise-linear coordinate system by backpropagation, regularization networks which take each data point as a nonlinear coordinate axis and learn the weights of the axes with regularization terms, and ensemble methods which approximate a nonlinear coordinate system by a piecewise-linear coordinate system. Ensemble methods have demonstrated their simplicity, robustness, and effectiveness in supervised learning, e.g. (Breiman, 1996, 2001; Dietterich, 2000; Freund and Schapire, 1995; Friedman et al, 2000), and are under fast development in unsupervised learning, e.g. (Dudoit and Fridlyand, 2003; Fern and Brodley, 2003; Fred and Jain, 2005; Strehl and Ghosh, 2003).

The present work *multilayer bootstrap network* (MBN) (Fig. 1) is a nonparametric model in a discrete space that approximates a data distribution with no model assumption and parameter tuning. It also generalizes data resampling (Efron, 1979; Efron and Tibshirani, 1993) and model ensembling methods to an unsupervised multilayer architecture.

A key feature of MBN is the training method of its  $k$ -centers clustering—random sampling, one-nearest neighbor optimization plus the binarization of its output, which outputs an invariant representation given any free parameters of a monotonous similarity metric in the original feature space. The similarity of two data points with the new representation is the number of their shared nearest centers in the clusterings. Although the discrete output of a single  $k$ -centers clustering loses much information comparing to a soft output in a continuous space, it can be proved from the analysis of the generalization error of supervised learning (Hastie et al, 2009) that the information loss of a group of  $k$ -centers clusterings is proportional to the correlation between the clusterings. The correlation is reduced by random sampling of data and random selection of features at each clustering, so as to the information loss.

Another key feature of MBN is a gradually narrowed multilayer architecture from bottom up. Different from supervised ensemble methods which aim to reduce the variance (and maybe also the bias) of predictions in a well-defined target space, a data distribution in a feature space may be highly variant. Hence, contrast to the popularity of learning one layer base learners in supervised learning, MBN builds a multilayer architecture to reduce larger and larger local variance of data layerwisely. It essentially builds an exponentially large number of hierarchical trees on the feature space (instead of directly on data points).

This paper is organized as follows. In Section 2, we describe MBN. In Section 3, we justify MBN theoretically. In Section 4, we introduce some related work. In Section 5, we study MBN empirically. In Section 6, we summarize our contributions. In Appendix, we analyze the theoretical computational complexity, propose a new algorithm *compressive MBN* to reduce the high prediction complexity of MBN, and report the results of the application of MBN to clustering on a number of benchmark data sets.

Table 1: Hyperparameters of MBN.

Parameter	Description
$L$	Number of layers.*
$\{k_l\}_{l=1}^L$	Number of centers per clustering where $l = 1, \dots, L$ is the index of the $l$ th layer. The clusterings in the same layer have the same number of centers.
$a$	Fraction of randomly selected dimensions (i.e., $\hat{d}$ ) over all dimensions (i.e., $d$ ) of input data.
$V$	Number of $k$ -centers clusterings per layer.

\* In practice,  $k_l$  decays with a factor of  $\delta$  (i.e.  $k_l = \delta k_{l-1}$  where  $l = 2, \dots, L$  and  $\delta = 0.5$ ), as a result, parameter  $L$  is determined by  $\{k_1, k_L, \delta\}$ .

## 2 Multilayer bootstrap networks

MBN contains multiple hidden layers and an output layer (Fig. 1). Each hidden layer is a group of mutually independent  $k$ -centers clusterings; each  $k$ -centers clustering has  $k$  output units, each of which indicates one cluster; the output units of all  $k$ -centers clusterings are concatenated as the input of their upper layer. The output layer is PCA. Parameter  $k$  should be as large as possible at the bottom layer and be smaller and smaller along with the increase of the number of layers until a predefined smallest  $k$  is reached.<sup>1</sup>

MBN is trained layer-by-layer. For training each layer given a  $d$ -dimensional input data set  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  either from the lower layer or from the original data space, we simply need to focus on training each  $k$ -centers clustering, which consists of the following three steps:

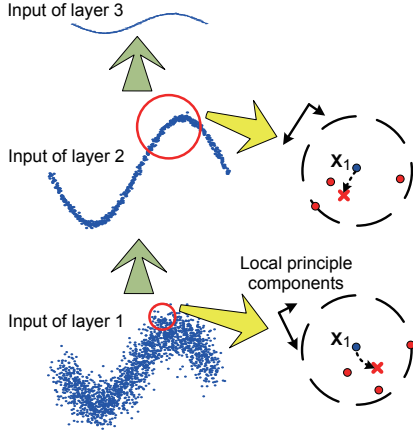
- **Random feature selection.** The first step randomly selects  $\hat{d}$  dimensions of  $\mathcal{X}$  ( $\hat{d} \leq d$ ) to form a subset of  $\mathcal{X}$ , denoted as  $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n\}$ .
- **Random sampling.** The second step randomly selects  $k$  data points from  $\hat{\mathcal{X}}$  as the  $k$  centers of the clustering, denoted as  $\{\mathbf{w}_1, \dots, \mathbf{w}_k\}$ .
- **Sparse representation learning.** The third step assigns each input data point  $\hat{\mathbf{x}}$  to one of the  $k$  clusters and outputs a  $k$ -dimensional indicator vector  $\mathbf{h} = [h_1, \dots, h_k]^T$ , where operator  $^T$  denotes the transpose of vector. For example, if  $\hat{\mathbf{x}}$  is assigned to the second cluster, then  $\mathbf{h} = [0, 1, 0, \dots, 0]^T$ . The assignment is calculated according to the similarities between  $\hat{\mathbf{x}}$  and the  $k$  centers, in terms of some predefined similarity measurement at the bottom layer, such as the squared Euclidean distance  $\arg \min_{i=1}^k \|\mathbf{w}_i - \hat{\mathbf{x}}\|^2$ , or in terms of  $\arg \max_{i=1}^k \mathbf{w}_i^T \hat{\mathbf{x}}$  at all other hidden layers.

MBN handles large-scale problems well. For training each layer, the time complexity is  $O(kV^2n)$ , and the storage complexity is  $O(2Vn)$ , where  $V$  is the number of clusterings per layer (see Table 1 for the notations of the hyperparameters, and Appendix A for a detailed analysis of the training complexity).

### 2.1 A typical hyperparameter setting

As shown in Section 5.3, MBN is robust to a wide range of hyperparameter settings. Here we introduce a typical setting which works well in terms of both effectiveness and efficiency.

<sup>1</sup> Parameter  $k$  cannot be too small. See Section 2.1 for a recommended smallest value of  $k$ .



**Fig. 2** Principle of MBN. The area in the red circle of a layer represents the local region of the data point  $\mathbf{x}_1$  at that layer, which is further amplified in a dashed circle on the right side. Each red point in the dashed circle is the closest center of a  $k$ -centers clustering to  $\mathbf{x}_1$ . The new representation of  $\mathbf{x}_1$  at the layer is located at the red cross in the dashed circle. The local principle components of the local region are shown in the upper and left corner of the dashed circle.

- **Setting hyperparameter  $k$ .** (i) For small-scale problems,  $k_1$  should be set around the middle size of the input data, i.e. around  $0.5n$ . For large-scale problems,  $k_1$  should be as large as possible. Suppose the largest  $k$  supported by hardware is  $k_{\max}$ , then  $k_1 = \min(0.7n, k_{\max})$ . (ii)  $k_l$  decays with a factor  $\delta$ , e.g.  $\delta = 0.5$ , with the increase of hidden layers. That is to say,  $k_l = 0.5k_{l-1}$ . (iii)  $k_L$  should be larger than the number of ground truth classes  $c$ . Typically,  $k_L \approx 1.5c$ . If  $c$  is unknown, we should make a rough guess on  $c$ .
- **Setting other hyperparameters.** Hyperparameter  $V$  should be at least larger than 100, typically  $V = 400$ . Hyperparameter  $a$  is fixed to 0.5. Hyperparameter  $L$  is determined by  $k$ .

### 3 Theoretical analysis

In this section, we first illustrate the geometrical interpretation of MBN, then present the importance of its binarization component (the third step of training a clustering), and at last analyze its theoretical base from the perspective of the bias-variance decomposition.

#### 3.1 Geometric interpretation

MBN has a simple geometric interpretation. As shown in Fig. 2, MBN first conducts piecewise-linear dimensionality reduction—a local PCA that gradually enlarges the area of a local region—implicitly in hidden layers, and then gets a low-dimensional feature explicitly by PCA. Specifically, each data point (e.g.,  $\mathbf{x}_1$  in Fig. 2) owns a local region supported by the centers of all clusterings that are closest to the data point. The centers define the local coordinate system. The new representation of the data point is the coordinates of the data point in the local coordinate system. If some other data points share the same local region, they will also be projected to the same coordinates, which means the small variances (i.e., small principle components) of this local region that are not covered by the local coordinate system will be discarded. It is easy to image that when  $k$  is smaller and smaller, the local

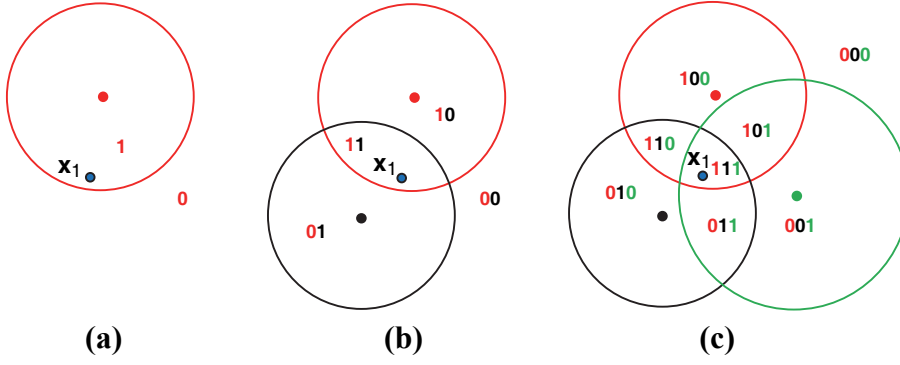


Fig. 3: Encoding the local region of a data point  $x_1$  (in blue color) by **a** one  $k$ -centers clustering, **b** two  $k$ -centers clusterings, and **c** three  $k$ -centers clusterings. The centers of the three  $k$ -centers clusterings are the points colored in red, black, and green respectively. The local region of a center is the area in the circle around the center, where the center and the edge of its local region have the same color.

region is gradually enlarged, making larger and larger relatively-unimportant local variances discarded.

It is important to keep parameter  $k$  of the  $k$ -centers clusterings in a layer the same. Otherwise, the density of centers between the clusterings are different, that is to say, the coordinate axes of a local region are built in different data spaces.

### 3.2 On the importance of binarization

MBN is an autoencoder. As shown in Fig. 3, the local region of a data point (e.g.  $x_1$ ) is partitioned to a number of disconnected fractions by the closest centers around the data point. Each fraction is assigned a unique binary code by *binarization* which is the step of the sparse representation learning. The new representation of the input data point is the binary code of the fraction where the data point is located.

The binarization is the base of MBN for dimensionality reduction. After the binarization, each small fraction with an arbitrary surface (such as shape and area) in the input space is shrunk to a single point in the discrete output space, represented as a unique binary code that encodes the position of the fraction. In other words, the variance of each fraction is reduced. The difference between two neighboring small fractions in the discrete output space is one bit. Note that a small fraction should not have to contain data points as shown in Fig. 3.

This encoding process learns an invariant representation of data from bottom up by first reducing the model complexity and then binarizing the data representation alternatively. Specifically, after reducing the variances of the small fractions in a layer by the binarization, it is able to make another locally linear assumption by reducing the model complexity in the successive upper layer. The model incorporates one or multiple binary codes into a new local small fraction; each new fraction is shrunk to a binary code again by the binarization. Suppose a binary output code at the  $L$ th layer is a merging of  $B_{L-1}$  binary codes ( $B_{L-1} \geq 1$ ) at the  $(L-1)$ th layer. Suppose each of the  $B_{L-1}$  binary codes is a merging of  $B_{L-2}$  binary codes at the  $(L-2)$ th layer and so on. The binary code at the  $L$ th layer is a shrinkage of a (maybe) highly-variant region that contains  $\prod_{l=1}^{L-1} B_l$  small fractions in the input space. As

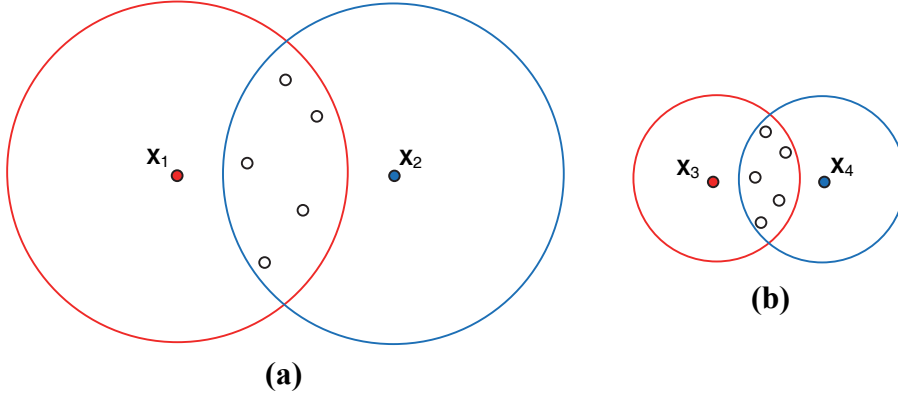


Fig. 4: Illustration of the similarity metric problem in the original feature space. **a** The similarity problem of two data points  $\mathbf{x}_1$  (in red color) and  $\mathbf{x}_2$  (in blue color) in a distribution  $\mathcal{P}_1$ . The local region of  $\mathbf{x}_1$  (or  $\mathbf{x}_2$ ) is the area in a colored circle that is centered at  $\mathbf{x}_1$  (or  $\mathbf{x}_2$ ). The small hollow points that lie in the cross area of the two local regions are the shared centers by  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . **b** The similarity problem of two data points  $\mathbf{x}_3$  and  $\mathbf{x}_4$  in a distribution  $\mathcal{P}_2$ .

a result, the data representation at the top layer is highly-invariant to the small variations of data, which is particularly useful for some applications, such as clustering. For example, for a ten class problem, a data point needs at most a 10-bits code to indicate its ground-truth label, leaving other information (i.e., small variations of data) unused, where the 10-bits code can be regarded as a highly-invariant representation of the data point. See Section 4.1 for more information on the hierarchical learning process.

The binarization is the key component for preventing MBN from hyperparameter tuning. Specifically, a major obstacle for adopting a soft representation is that the similarity metric may not fit the true data distribution. For example, as shown in Fig. 4, the similarity between the data points that are far apart in a distribution with a large variance (e.g.,  $\mathbf{x}_1$  and  $\mathbf{x}_2$  in a distribution  $\mathcal{P}_1$ ) might be the same as the similarity between the data points that are close to each other in a distribution with a small variance (e.g.,  $\mathbf{x}_3$  and  $\mathbf{x}_4$  in a distribution  $\mathcal{P}_2$ ). If we use the Euclidean distance as the similarity measurement, then the similarity between  $\mathbf{x}_1$  and  $\mathbf{x}_2$  is smaller than that between  $\mathbf{x}_3$  and  $\mathbf{x}_4$ , which is not true. To overcome this problem, one common method is to fit a predefined model, such as GMM, to the data distribution (see Section 4.2); another common method is to introduce a tunable similarity metric, such as the Gaussian radial basis function (RBF) (see Section 4.3).

MBN provides an adaptive similarity metric. (i) The uniform resampling, nearest neighbor optimization, plus binarization provide a distributed nonparametric method for estimating the density of data. For example, in Fig. 4, if the local region in  $\mathcal{P}_1$  is partitioned in the same way as the local region in  $\mathcal{P}_2$ , and if the only difference between them is that the local region in  $\mathcal{P}_1$  is an amplification of the local region in  $\mathcal{P}_2$ , then the surfaces of the two local regions are the same in the discrete output space. (ii) After binarizing the data representation, the similarity between two data points are measured by the number of the nearest centers they share in common. For example,  $\mathbf{x}_1$  and  $\mathbf{x}_2$  share 5 common nearest centers, and  $\mathbf{x}_3$  and  $\mathbf{x}_4$  also share 5 common nearest centers, so that the similarity between  $\mathbf{x}_1$  and  $\mathbf{x}_2$  in

$\mathcal{P}_1$  equals to the similarity between  $\mathbf{x}_3$  and  $\mathbf{x}_4$  in  $\mathcal{P}_2$ . Moreover, the binarization makes MBN insensitive to outliers. It also reduces the time and storage complexities.

The binarization does not lose information as much as it appears. Specifically, for a single  $k$ -centers clustering, a binarized output indeed loses much more information than a soft output. However, as shown in Fig. 3c,  $V$   $k$ -centers clusterings can partition a local region to  $2^V$  disconnected small fractions at the maximum. It is easy to image that, for any local region, when  $V$  is increasing and the diversity between the centers is still reserved, an ensemble of  $k$ -centers clusterings approximates the true data distribution. Because the diversity is important, we have used two randomized steps to enlarge it. See Section 3.3 for a theoretical analysis on this problem.

### 3.3 Theoretical base

In this subsection, we analyze the effectiveness of MBN by theoretically evaluating how accurate the output sparse representation can be after data resampling and model ensemble, comparing to a single  $k$ -centers clustering. Suppose the correlation coefficient between two random samples (i.e.,  $\{\mathbf{w}_{v_1,i}\}_{i=1}^k$  and  $\{\mathbf{w}_{v_2,j}\}_{j=1}^k$ ,  $\forall v_1, v_2 = 1, \dots, V$  and  $v_1 \neq v_2$ ) is  $\rho$ ,  $0 \leq \rho \leq 1$ . Each random sample formulates a  $k$ -centers clustering and produces a sparse representation of data. Suppose the variance of the sparse representation produced by a single random sample is  $\sigma_{\text{single}}^2$ , and the variance of the sparse representation produced by an ensemble of random samples is  $\sigma_{\text{ensemble}}^2$ .  $\sigma_{\text{single}}^2$  and  $\sigma_{\text{ensemble}}^2$  have the following relation:

**Theorem 1** When  $\rho$  is reduced from 1 to 0,  $\sigma_{\text{ensemble}}^2$  is reduced from  $\sigma_{\text{single}}^2$  to  $\sigma_{\text{single}}^2/V$  accordingly.

*Proof* We prove Theorem 1 by first transferring MBN to a supervised regression problem and then using the bias-variance decomposition of the mean squared error to get the bias and variance components of a relaxed version of MBN. The detail is as follows.

We focus on analyzing a given point  $\mathbf{x}$ , and assume that the *true* local coordinate of  $\mathbf{x}$  is  $\mathbf{s}$  which is an invariant point around  $\mathbf{x}$  and usually found when the density of the nearest centers around  $\mathbf{x}$  goes to infinity (i.e.  $n \rightarrow \infty, V \rightarrow \infty$ ,  $\{\mathbf{w}_v\}_{v=1}^V$  are i.i.d., and parameter  $k$  is unchanged). We also suppose that  $\mathbf{x}$  is projected to  $\hat{\mathbf{s}}$  when given a finite number of nearest centers  $\{\mathbf{w}_v\}_{v=1}^V$ . Hence, the effectiveness of MBN can be evaluated by the robustness of the estimate  $\hat{\mathbf{s}}$  to the truth  $\mathbf{s}$ . Note that  $\mathbf{s}$  is used as an invariant reference point for studying  $\hat{\mathbf{s}}$ .

As analyzed in Section 3.2, after the binarization, the similarity between  $\hat{\mathbf{s}}$  and  $\mathbf{s}$  is evaluated by the proportion of the nearest centers they share in common to  $\{\mathbf{w}_v\}_{v=1}^V$ . For each nearest center  $\mathbf{w}_v$ , the problem can be formulated as a two-class classification problem: if both  $\hat{\mathbf{s}}$  and  $\mathbf{s}$  are in the local region of  $\mathbf{w}_v$ , then  $\hat{f}_v(\hat{\mathbf{s}}, \mathbf{s}) = 1$ ; otherwise  $\hat{f}_v(\hat{\mathbf{s}}, \mathbf{s}) = 0$ , where  $\hat{f}_v(\cdot)$  is the classifier. The overall similarity between  $\hat{\mathbf{s}}$  and  $\mathbf{s}$  is calculated by  $\hat{f}_\Sigma(\hat{\mathbf{s}}, \mathbf{s}) = \frac{1}{V} \sum_{v=1}^V \hat{f}_v(\hat{\mathbf{s}}, \mathbf{s})$ . Because the points  $\mathbf{x}$ ,  $\hat{\mathbf{s}}$ , and  $\mathbf{w}_v$  are always in the local region of  $\mathbf{w}_v$ ,  $\hat{\mathbf{s}}$  can be omitted from  $\hat{f}_v(\hat{\mathbf{s}}, \mathbf{s})$  when given  $\mathbf{w}_v$ , which simplifies the calculation of the similarity between  $\hat{\mathbf{s}}$  and  $\mathbf{s}$  to the following equation:

$$\hat{f}_\Sigma(\mathbf{s}) = \frac{1}{V} \sum_{v=1}^V \hat{f}_v(\mathbf{s}) \quad (1)$$

where

$$\hat{f}_v(\mathbf{s}) = \begin{cases} 1, & \text{if } \mathbf{s} \text{ is in the local region of } \mathbf{w}_v \\ 0, & \text{otherwise} \end{cases} \quad \forall v = 1, \dots, V. \quad (2)$$

Eq. (1) makes us clear that the similarity between  $\hat{\mathbf{s}}$  and  $\mathbf{s}$  is essentially the voting result of an ensemble of one-nearest-neighbor classifiers. However, unlike supervised classification, the output of Eq. (2) is decided not only by the distance between  $\mathbf{w}_v$  and  $\mathbf{s}$  but also by the local area of  $\mathbf{w}_v$  which makes problem (1) unnecessarily overcomplicated.

To prevent this overcomplicated problem, we relax problem (2) to a problem in the Euclidean space:

$$\hat{f}_v(\mathbf{s}) = \exp\left(-\frac{\|\mathbf{s} - \mathbf{w}_v\|^2}{2\sigma^2}\right), \quad \forall v = 1, \dots, V \quad (3)$$

where  $\sigma$  is a parameter that is strongly related to the local area of  $\mathbf{w}_v$ , and  $\|\mathbf{s} - \mathbf{w}_v\| \sim \mathcal{N}(0, \sigma_{\text{input}}^2)$  where  $\mathcal{N}(0, \sigma_{\text{input}}^2)$  denotes a univariate normal distribution with zero mean and a variance of  $\sigma_{\text{input}}^2$  (see Theorem 2 for a further discussion of the relaxation of problem (2)).<sup>2</sup> Because  $\mathbf{s}$  has a very high probability to be assigned to 1 when  $\|\mathbf{s} - \mathbf{w}_v\|^2 < \sigma_{\text{input}}^2$ , it is clear that  $\sigma^2 \geq \sigma_{\text{input}}^2$ .

Due to this relaxation, we are able to reformulate problem (1) to a supervised regression problem where  $\mathbf{s}$  is the input and  $f(\mathbf{s}) = 1$  is the *truth*. It is known that the mean squared error of a regression problem can be decomposed to the summation of a squared bias component and a variance component (Hastie et al, 2009):

$$\begin{aligned} E\left((f(\mathbf{s}) - \hat{f}(\mathbf{s}))^2\right) &= (f(\mathbf{s}) - E(\hat{f}(\mathbf{s})))^2 + E\left((\hat{f}(\mathbf{s}) - E(\hat{f}(\mathbf{s})))^2\right) \\ &= \text{Bias}^2(\hat{f}(\mathbf{s})) + \text{Var}(\hat{f}(\mathbf{s})) \end{aligned} \quad (4)$$

where  $E(\cdot)$  is the expectation of a random variable, and  $\hat{f}(\mathbf{s})$  is an estimate of  $f(\mathbf{s})$ . Given

$$E(\hat{f}_v(\mathbf{s})) = \frac{\sigma}{\sqrt{\sigma_{\text{input}}^2 + \sigma^2}}, \quad \forall v = 1, \dots, V \quad (5)$$

$$E\left((\hat{f}_v(\mathbf{s}))^2\right) = \frac{\sigma}{\sqrt{2\sigma_{\text{input}}^2 + \sigma^2}} = \frac{\sigma^2}{\sqrt{(\sigma_{\text{input}}^2 + \sigma^2)^2 - \sigma_{\text{input}}^4}}, \quad \forall v = 1, \dots, V \quad (6)$$

$$\begin{aligned} E(\hat{f}_{v_1}(\mathbf{s})\hat{f}_{v_2}(\mathbf{s})) &= \frac{\sigma^2}{\sqrt{(1-\rho^2)\sigma_{\text{input}}^4 + 2\sigma_{\text{input}}^2\sigma^2 + \sigma^4}} = \frac{\sigma^2}{\sqrt{(\sigma_{\text{input}}^2 + \sigma^2)^2 - \rho^2\sigma_{\text{input}}^4}}, \\ &\quad \forall v_1 = 1, \dots, V, \quad \forall v_2 = 1, \dots, V, \quad v_1 \neq v_2 \end{aligned} \quad (7)$$

we derive

$$\text{Bias}^2(\hat{f}_v(\mathbf{s})) = \left(1 - \frac{\sigma}{\sqrt{\sigma_{\text{input}}^2 + \sigma^2}}\right)^2 \quad (8)$$

<sup>2</sup> Given Eq. (3), Eq. (1) appears to be the marginal density of a GMM whose components have the same variance and prior probability. We clear up their differences as follows. The components of the GMM, which are localized at different regions with soft boundaries between them, represent different classes. However, the components in Eq. (1) represent Gaussian-like voting functions for the same class, and the centers of the components come from a single Gaussian distribution. An example of Eq. (1) is that two centers that have the same variance may have the same coordinate.



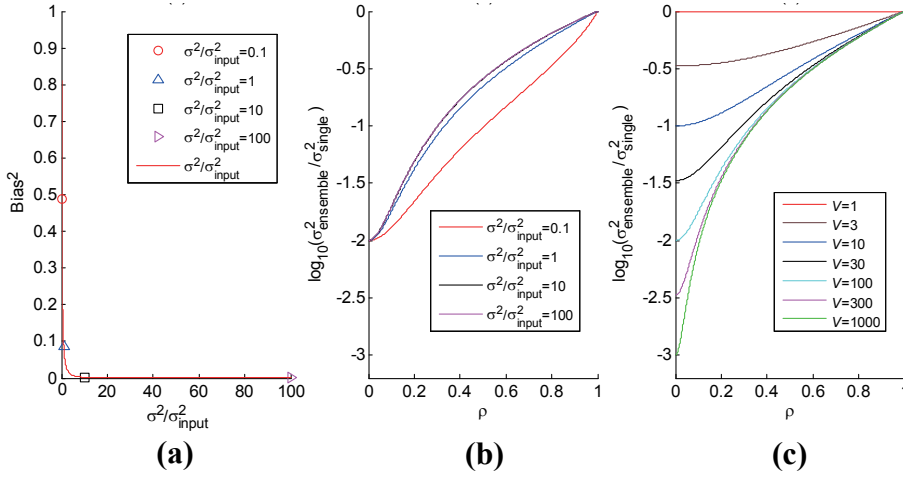


Fig. 5: Analysis of the bias and variance components of MBN. Note that figure **c** is drawn with  $\sigma^2/\sigma_{\text{input}}^2 = 1$ . The curves with  $\sigma^2/\sigma_{\text{input}}^2 > 1$  are similar with the curves in figure **c**.

$$\sigma_{\text{single}}^2 = \text{Var}(\hat{f}_v(\mathbf{s})) = \frac{\sigma^2}{\sqrt{(\sigma_{\text{input}}^2 + \sigma^2)^2 - \sigma_{\text{input}}^4}} - \frac{\sigma^2}{\sigma_{\text{input}}^2 + \sigma^2} \quad (9)$$

and

$$\text{Bias}^2(\hat{f}_{\Sigma}(\mathbf{s})) = \left(1 - \frac{\sigma}{\sqrt{\sigma_{\text{input}}^2 + \sigma^2}}\right)^2 \quad (10)$$

$$\begin{aligned} \sigma_{\text{ensemble}}^2 = \text{Var}(\hat{f}_{\Sigma}(\mathbf{s})) &= \frac{1}{V} \left( \frac{\sigma^2}{\sqrt{(\sigma_{\text{input}}^2 + \sigma^2)^2 - \sigma_{\text{input}}^4}} - \frac{\sigma^2}{\sqrt{(\sigma_{\text{input}}^2 + \sigma^2)^2 - \rho^2 \sigma_{\text{input}}^4}} \right) \\ &+ \left( \frac{\sigma^2}{\sqrt{(\sigma_{\text{input}}^2 + \sigma^2)^2 - \rho^2 \sigma_{\text{input}}^4}} - \frac{\sigma^2}{\sigma_{\text{input}}^2 + \sigma^2} \right). \end{aligned} \quad (11)$$

We can see that  $\text{Bias}^2(\hat{f}_v(\mathbf{s})) = \text{Bias}^2(\hat{f}_{\Sigma}(\mathbf{s}))$  and  $\sigma_{\text{ensemble}}^2 \in [\sigma_{\text{single}}^2/V, \sigma_{\text{single}}^2]$ . Function  $\sigma_{\text{ensemble}}^2/\sigma_{\text{single}}^2$  is a monotonic function that reaches  $1/V$  with  $\rho = 0$  and reaches 1 with  $\rho = 1$ . Theorem 1 is proved.  $\square$

Theorem 1 and its proof show that  $\mathbf{x}$  is likely in the same small fraction as or around its true local coordinate, and the small fraction where  $\mathbf{x}$  is located has a small variance, which in turn supports the effectiveness of MBN.

We further analyze the bias and variance components in Fig. 5. The condition for the meaningfulness of Eq. (3) is likely  $\sigma^2/\sigma_{\text{input}}^2 \geq 1$ . For fully understanding the bias and variance functions, we also draw the results of the unlikely scenario  $\sigma^2/\sigma_{\text{input}}^2 < 1$  without a further analysis. Fig. 5a shows that  $\text{Bias}^2(\hat{f}_\Sigma(\mathbf{s}))$  with  $\sigma^2/\sigma_{\text{input}}^2 \geq 1$  is low. Fig. 5b shows that  $\log_{10}(\sigma_{\text{ensemble}}^2/\sigma_{\text{single}}^2)$  is monotonic with respect to  $\rho$ . Fig. 5c shows that  $\log_{10}(\sigma_{\text{ensemble}}^2/\sigma_{\text{single}}^2)$  is gradually getting smaller at  $\rho = 0$  when  $V$  is getting larger. Fig. 5c also shows that the effectiveness of the model ensembling has a strong dependency on  $\rho$ : A model with a larger  $V$  needs a smaller  $\rho$  to apparently show its advantage over a model with a smaller  $V$ ; simply enlarging  $V$  without de-correlating the base clusterings results in limited performance improvement, which is our motivation for the development of the first step of training the base clusterings. In practice, when  $\mathcal{X}$  is given, enlarging  $k$  and decreasing  $\rho$  is a pair of contradictory factors.

Note that we cannot use Eq. (3) to score  $\mathbf{x}$  directly, since the assumption that  $\{\mathbf{w}_v\}_{v=1}^V$  is a Gaussian distribution centered at  $\mathbf{x}$  is too strong in practice. A counter-example is shown in Fig. 2. Even if the assumption happens to be correct, we have to estimate the local variance of each data point, which is impractical. As a word, Eq. (3) is no more than a relaxation of Eq. (2) for facilitating our theoretical analysis at a single point.

## 4 Relationship with other methods

MBN is related to many machine learning methods. Here we introduce its connection to hierarchical clustering, mixture of experts, RBF networks, bootstrap methods, clustering ensemble, and sparse coding.

### 4.1 Hierarchical clustering

Hierarchical clustering, either agglomerative or divisive, embeds a tree into data. Each leaf node of the tree contains a single data point; the root node contains all data points; and a father node is a merging of the data points of its child nodes. Agglomerative clustering, which is more common than divisive clustering, initially takes each data point as a single cluster and then sequentially merges the two clusters that are more similar to each other than any other pair of clusters.

Agglomerative clustering is effective in capturing the local patterns of a data distribution. However, it suffers from the following weaknesses. First, the leaf nodes of an agglomerative clustering partition the input space to only  $n$  local regions; and the merging of two nodes, which discards the detailed local information, is rough. Due to this rough partition of the input space, the tree structure exposes its weakness that if a child node is merged to a wrong father node caused by some “bad” data points in the child node, then all “good” data points in the child node will have no chance to be corrected during the remaining clustering process. Moreover, it needs to recalculate the similarity matrix of the clusters after each merging, so that it generally has a computational complexity of  $O(n^3)$  and a storage complexity of  $O(n^2)$  which makes it not suitable for large-scale problems.

MBN does not suffer from these weaknesses. Specifically, MBN builds trees implicitly: Each binary output code at the top layer is the root node of a tree whose leaf nodes are a number of small fractions in the input space. However, because there are as many as an order of  $k_L 2^V$  root nodes (i.e. the trees) which are far more than the data points in practice, any two

data points have a very small probability to be merged into a single root node. That is to say, the output of MBN is the representation of data but not data clusters, though MBN happens to merge the data points who share exactly the same nearest centers. Moreover, MBN has linear time and storage complexities, so that it can deal with large-scale problems.

#### 4.2 Mixture of experts

A mixture of experts assigns a base model (called an expert) to each local region. Its parameters are estimated by maximizing the likelihood  $p(\mathbf{x}) = \sum_i \alpha_i g_i(\mathbf{x})$  where  $g_i(\mathbf{x})$  is the  $i$ th expert and  $\{\alpha_i\}_i$  is a gating function with  $\sum_i \alpha_i = 1$  and  $\alpha_i > 0$ . The maximum likelihood estimate is usually found by the expectation-maximization algorithm. Two common mixtures of experts are  $k$ -means clustering and GMM.

Mixture of experts is easily understood and widely used. However, it may make an inaccurate model assumption. For example, GMM makes a strong assumption that an underlying data distribution can be estimated by a fixed number of Gaussian models, which may be inaccurate, particularly for a highly-nonlinear data distribution. Moreover, mixture of experts generates a localized representation by assigning each local region an expert, which can be quite costly and even infeasible for a rather complicated data distribution. At last, training a mixture of experts by expectation-maximization is computationally expensive and suffers from local minima, particularly when the number of mixtures is large.

MBN overcomes these weaknesses. First, MBN does not make specific model assumptions. Because the  $k$  centers of an expert (i.e., a base clustering) is sampled directly from data, the expert can produce a reasonable approximation of the true data distribution as if  $k$  is large enough. Second, MBN generates a distributed representation which can be exponentially more effective than a localized representation. As shown in Fig. 3, a layer of MBN needs only  $kV$  hidden units to partition the input space to an order of  $k2^V$  disconnected fractions, while a hard  $k$ -means clustering needs probably  $k2^V$  centers to obtain the same representation. Note that a  $k$ -means clustering can at most partition the input into  $n$  disconnected fractions. Third, MBN does not need the expectation-maximization algorithm to optimize the centers of the experts, since that, when  $k$  is large enough, optimizing the centers is statistically meaningless.

MBN demonstrates the advantage of learning a hierarchical representation over learning a single layer of representation. Specifically, as presented in Section 3.2, the binary code of a data point at the top layer represents a large number of (e.g.  $\prod_{l=1}^{L-1} B_l$ ) small fractions in the input space which can be an arbitrary distribution. However, the code of a data point produced by a mixture of experts represents a local region with a fixed distribution, e.g. Gaussian distribution.

#### 4.3 Radial basis function networks

Function  $\phi(\|\mathbf{x}\|)$  is a radial function if its value depends only on the distance (usually defined on a Euclidean space) between  $\mathbf{x}$  and the origin. It is a monotonic function and converges to zero with increasing  $\|\mathbf{x}\|$ . If we use a data point  $\mathbf{w}_i$  as the center instead of the origin, function  $\phi(\|\mathbf{x} - \mathbf{w}_i\|)$  is called a RBF. RBF usually contains free parameters, such as the Gaussian RBF  $\phi(\|\mathbf{x} - \mathbf{w}_i\|) = \exp(-\gamma\|\mathbf{x} - \mathbf{w}_i\|^2)$  where  $\gamma$  is a free parameter. It is known that the performance of a RBF network relies on its free parameters.

MBN can be viewed as an unsupervised multilayer RBF network, but its performance does not rely on free parameters of RBF. Given an input  $\mathbf{x}$ , a  $k$ -centers clustering learns a  $k$ -dimensional sparse vector  $\mathbf{h}$  from a RBF representation  $[\phi(\|\mathbf{x} - \mathbf{w}_1\|), \dots, \phi(\|\mathbf{x} - \mathbf{w}_k\|)]^T$ :

$$h_m = \begin{cases} 1, & \text{if } m = \arg \max_{i=1}^k \phi(\|\mathbf{x} - \mathbf{w}_i\|) \\ 0, & \text{otherwise} \end{cases} \quad \forall m = 1, \dots, k. \quad (12)$$

Because  $\phi(\|\mathbf{x} - \mathbf{w}_i\|)$  is a monotonic function,  $\mathbf{h}$  is irrelevant to the choice of RBF thanks to the binarization in Eq. (12), which makes the similarity metric at the bottom layer of MBN any reasonable one, such as the common squared Euclidean distance  $m = \arg \min_{i=1}^k \|\mathbf{x} - \mathbf{w}_i\|^2$ . If a hidden layer is not at the bottom, then  $\arg \min_{i=1}^k \|\mathbf{x} - \mathbf{w}_i\|^2 = \arg \max_{i=1}^k \mathbf{x}^T \mathbf{w}_i$ , where  $\mathbf{x}^T \mathbf{w}_i$  is computationally more efficient than  $\|\mathbf{x} - \mathbf{w}_i\|^2$ .

#### 4.4 Bootstrap methods

Bootstrap resampling (Efron, 1979; Efron and Tibshirani, 1993) has been applied successfully to machine learning, where the phrase *bootstrap* comes from the proverb *to pull oneself up by one's bootstrap*. Given a data set  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , a bootstrap sample is obtained by randomly sampling  $\mathcal{X}$   $n$  times, *with replacement*. For example, a bootstrap sample of  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_6\}$  may be  $\mathcal{X}^* = \{\mathbf{x}_1^* = \mathbf{x}_2, \mathbf{x}_2^* = \mathbf{x}_5, \mathbf{x}_3^* = \mathbf{x}_1, \mathbf{x}_4^* = \mathbf{x}_2, \mathbf{x}_5^* = \mathbf{x}_4, \mathbf{x}_6^* = \mathbf{x}_1\}$ . Give an estimator  $\theta = f(\mathcal{X})$  and bootstrap samples  $\{\mathcal{X}_v^*\}_{v=1}^V$ , one can use bootstrap replicates  $\{\theta_v^* = f(\mathcal{X}_v^*)\}_{v=1}^V$  to make inference. One of the key ideas of supervised bootstrap methods in machine learning is that bootstrap resampling is an effective way of reducing the variance of predictions (Hastie et al, 2009). Bootstrap aggregation (*bagging*) (Breiman, 1996) builds a regression tree on each bootstrap sample (originally without replacement) and aggregates the prediction results of the trees, which reduces the prediction variance from  $\sigma^2$  by a single tree to  $\rho\sigma^2 + \frac{1-\rho}{V}\sigma^2$  where  $0 \leq \rho \leq 1$  is the correlation coefficient between the trees. Random forests (Breiman, 2001) further reduce  $\rho$  by the random selection of features at each node.

MBN has several differences from the above supervised bootstrap methods. First, MBN uses random subsampling *without replacement*. The reason why we do not adopt the standard bootstrap resampling is that the resampling in MBN is used to build local coordinate systems, hence, if a data point is sampled multiple times, the duplicated data points are still viewed as a single coordinate axis. Moreover, bootstrap resampling will make parameter  $k$  of the  $k$ -centers clusterings in a layer not a constant, which is an undesired behavior for MBN. Second, although the variance components of MBN and the supervised bootstrap methods share a similar advantage, MBN aims to reduce the small local variances of data instead of reducing the variance of predictions. Third, MBN is built from bottom up stacking instead of top-down splitting. Fourth, MBN builds a vast number of local trees implicitly in the feature space instead of building a handful global trees directly on the input data points explicitly. However, MBN was motivated from and shares many common properties with the bootstrap methods, such as building each base clustering from a random sample of the input and decorrelating the base clusterings by the random selection of features at each base clustering, hence, we adopted the phrase “bootstrap” in MBN and clarify its usage here for preventing confusion.

#### 4.5 Clustering ensemble

Clustering ensemble (Dudoit and Fridlyand, 2003; Fern and Brodley, 2003; Fred and Jain, 2005; Strehl and Ghosh, 2003; Vega-Pons and Ruiz-Shulcloper, 2011) is a clustering technique that uses a *consensus function* to aggregate the clustering results of a set of mutually-independent base clusterings. It constructs diverse base clusterings in the four ways summarized in (Dietterich, 2000). For example, the algorithm in (Dudoit and Fridlyand, 2003) builds each base clustering from a bootstrap sample; the algorithm in (Fern and Brodley, 2003) introduces randomization into the feature space by random projection; the algorithm in (Fred and Jain, 2005) constructs diverse  $k$ -means clusterings via the local-minima property of  $k$ -means; the algorithm in (Strehl and Ghosh, 2003) uses different types of base clusterings together, such as partition clusterings and density clusterings. Besides, many clustering ensemble techniques focused on designing the consensus function (Strehl and Ghosh, 2003). See (Vega-Pons and Ruiz-Shulcloper, 2011) for an overview of clustering ensemble.

Clustering ensemble reduces the small variances of data by one layer of base clusterings, leaving the ensembling technique as a method of reducing the variance of the clustering result. A key problem of the base clustering algorithms is that they reduce the small variances of data too aggressively. For example, partition clusterings, such as  $k$ -means, use the gravity center of a cluster as the representation of all data points that fall into the cluster, and discard all nonlinear variations within the cluster. Hierarchical clusterings iteratively merge (or divide) clusters and use the gravity center of a cluster to represent all data points that fall into the cluster. Density clusterings use a data point with a high local density to represent all points that fall into the local dense region. Although probabilistic clusterings, such as Gaussian mixture model, represent the data points that are belonged to the same cluster by a probabilistic distribution around a cluster center, they have to make a model assumption on the underlying data distribution and estimate the variance of the model additionally. After an aggressive variance reduction of data, the clustering result of each base clustering may not be accurate enough, particularly when the number of the output clusters is set to the ground-truth number which is true to the literature of clustering ensemble. An exceptional clustering ensemble method is (Fred and Jain, 2005), in which each base  $k$ -means clustering produces a subclass partition by assigning the parameter  $k$  to a random value that is slightly larger than the ground-truth cluster number.

MBN uses data resampling plus model ensembling to reduce the small local variances of data gradually in a multilayer architecture, while clustering ensemble uses model ensembling techniques to reduce the variance of the clustering result. To integrate the merits of MBN and clustering ensemble, we may apply the low-dimensional output of MBN to a clustering ensemble for a stable clustering result, if MBN is applied for clustering.

#### 4.6 Sparse coding

Given a learned dictionary  $\mathbf{W}$ , sparse coding typically aims to solve  $\min_{\mathbf{h}_i} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{W}\mathbf{h}_i\|_2^2 + \lambda \|\mathbf{h}_i\|_1$ , where  $\|\cdot\|_q$  represents  $\ell_q$ -norm,  $\mathbf{h}_i$  is the sparse code of the data point  $\mathbf{x}_i$ , and  $\lambda$  is a hyperparameter controlling the sparsity of  $\mathbf{h}_i$ . Each column of  $\mathbf{W}$  is called a basis vector. We may also view  $\lambda$  as a hyperparameter that controls the number of clusterings. Specifically, if we set  $\lambda = 0$ , it is likely that  $\mathbf{h}_i$  contains only one nonzero element. Intuitively, we can understand it as that we use only one clustering to learn a sparse code. A good value of  $\lambda$  can make a small part of the elements of  $\mathbf{h}_i$  nonzero. This choice approximates to the method of

partitioning the dictionary to several (probably overlapped) subsets and then grouping the basis vectors in each subset to a base clustering.

Empirically, Coates and Ng (2011) conducted a broad experimental comparison on sparse coding, and observed that using random sampling to form a dictionary and using *soft threshold* to extract sparse features can be highly competitive to complicated sparse coding methods. Because the method in (Coates and Ng, 2011) can be regarded as a single  $k$ -centers clustering that uses a few nearest centers of a data point as the nonzero elements of the output sparse code of the data point. As a mixture of experts, the method in (Coates and Ng, 2011) is less effective than the building block of MBN theoretically.

To summarize, we show the relationship between sparse coding and MBN formally:

**Theorem 2** *The  $\ell_1$ -norm sparse coding is a convex relaxation of the building block of MBN when given the same dictionary.*

*Proof* Each layer of MBN maximizes the likelihood of the following equation:

$$p(\mathbf{x}) = \prod_{v=1}^V g_v(\mathbf{x}) \quad (13)$$

where  $g_v(\mathbf{x})$  is a  $k$ -means clustering with the squared error as the similarity metric:

$$g_v(\mathbf{x}) = \mathcal{MN}(\mathbf{x}; \mathbf{W}_v \mathbf{h}_v, \sigma^2 \mathbf{I}) \quad (14)$$

subject to  $\mathbf{h}_v$  is a one-hot code

where  $\mathcal{MN}$  denotes the multivariate normal distribution,  $\mathbf{W}_v = [\mathbf{w}_{v,1}, \dots, \mathbf{w}_{v,k}]$  is the weight matrix whose columns are centers,  $\mathbf{I}$  is the identity matrix, and  $\sigma \rightarrow 0$ .

Given a data set  $\{\mathbf{x}_i\}_{i=1}^n$ . We assume that  $\{\mathbf{W}_v\}_{v=1}^V$  are fixed, and take the negative logarithm of Eq. (13):

$$\min_{\{\{\mathbf{h}_{v,i}\}_{i=1}^n\}_{v=1}^V} \sum_{v=1}^V \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{W}_v \mathbf{h}_{v,i}\|_2^2, \quad (15)$$

subject to  $\mathbf{h}_{v,i}$  is a one-hot code.

If we denote  $\mathbf{W} = [\mathbf{W}_1, \dots, \mathbf{W}_V]$  and further complement the head and tail of  $\mathbf{h}_{v,i}$  with multiple zeros, denoted as  $\mathbf{h}'_{v,i}$ , such that  $\mathbf{W} \mathbf{h}'_{v,i} = \mathbf{W}_v \mathbf{h}_{v,i}$ , we can rewrite Eq. (15) to the following equivalent problem:

$$\min_{\{\{\mathbf{h}'_{v,i}\}_{i=1}^n\}_{v=1}^V} \sum_{v=1}^V \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{W} \mathbf{h}'_{v,i}\|_2^2, \quad (16)$$

subject to  $\mathbf{h}'_{v,i}$  is a one-hot code.

It is an integer optimization problem that has an integer matrix variable  $\mathbf{H}'_v = [\mathbf{h}'_{v,1}, \dots, \mathbf{h}'_{v,n}]$ . Suppose there are totally  $|\mathcal{H}'_v|$  possible solutions of  $\mathbf{H}'_v$ , denoted as  $\mathbf{H}'_{v,1}, \dots, \mathbf{H}'_{v,|\mathcal{H}'_v|}$ , we first relax Eq. (16) to a convex optimization problem by constructing a *convex hull* (Boyd and Vandenberghe, 2004) on  $\mathbf{H}'_v$ :

$$\min_{\left\{\{\mu_{v,k}\}_{k=1}^{|\mathcal{H}'_v|}\right\}_{v=1}^V} \sum_{v=1}^V \sum_{i=1}^n \left\| \mathbf{x}_i - \mathbf{W} \left( \sum_{k=1}^{|\mathcal{H}'_v|} \mu_{v,k} \mathbf{h}'_{v,k,i} \right) \right\|_2^2 \quad (17)$$

subject to  $0 \leq \mu_{v,k} \leq 1, \sum_{k=1}^{|\mathcal{H}'_v|} \mu_{v,k} = 1, \quad \forall v = 1, \dots, V.$

Because Eq. (17) is a convex optimization problem, according to Jensen’s inequality, the following problem learns a lower bound of Eq. (17):

$$\begin{aligned} \min_{\left\{\{\mu_{v,k}\}_{k=1}^{|\mathcal{H}'_v|}\right\}_{v=1}^V} & V \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{W} \mathbf{h}_i''\|_2^2 \\ \text{subject to} & 0 \leq \mu_{v,k} \leq 1, \sum_{k=1}^{|\mathcal{H}'_v|} \mu_{v,k} = 1, \quad \forall v = 1, \dots, V \end{aligned} \quad (18)$$

where  $\mathbf{h}_i'' = \frac{1}{V} \sum_{v=1}^V \sum_{k=1}^{|\mathcal{H}'_v|} \mu_{v,k} \mathbf{h}'_{v,k,i}$  with  $\mu_{v,k}$  as a variable.

Recalling the definition of sparse coding given a fixed dictionary  $\mathbf{W}$ , we observe that Eq. (18) is a special form of sparse coding with more strict constraints on the format of sparsity.

Therefore, given the same dictionary  $\mathbf{W}$ , each layer of MBN is a distributed sparse coding that is lower bounded by the common  $\ell_1$ -norm sparse coding. When we discard the expectation-maximization optimization of each  $k$ -means clustering (i.e., dictionary learning) but only preserve the default initialization method – random sampling, Eq. (13) becomes the building block of MBN. Given the same dictionary, the  $\ell_1$ -norm-regularized sparse coding is a convex relaxation of the building block of MBN. Theorem 2 is proved.  $\square$

## 5 Empirical evaluation of multilayer bootstrap networks

In this section, we apply the low dimensional output of MBN to the tasks of visualizing and clustering the MNIST handwritten digits (Lecun et al, 2004) and retrieving the Reuters newswire stories (Lewis et al, 2004). When a data set is small-scale, we use the linear-kernel-based kernel PCA (Canu et al, 2005; Schölkopf et al, 1998) as the PCA toolbox of MBN. When a data set is middle- or large-scale, we use the expectation-maximization PCA (EM-PCA) (Roweis, 1998) as the PCA toolbox.<sup>3</sup> Kernel PCA can handle a data set that the dimension of the data is larger than the size of the data set, but its time and storage complexities scale squarely with the size of the data set. EM-PCA can handle large-scale and high-dimensional problems efficiently, but it suffers from local minima. When we evaluated the *training time*, all comparison methods were run with a *one-core* personal computer with 8 GB memory. Only the CPU time consumed on dimensionality reduction was recorded.

### 5.1 Data visualization and clustering

The data set of the MNIST digits (Lecun et al, 2004) contains 10 handwritten integer digits ranging from 0 to 9. It consists of 60,000 training images and 10,000 test images. Each image has 784 dimensions. The largest and smallest values of MNIST are 255 and 0 respectively. We normalized each image to the range  $[0, 1]$  by dividing each entry of the image by 255. Note that MBN does not have to use normalized data. The normalization is for the methods that have to use normalized data. It also helps preventing numerical problems of computer.

For the task of clustering, we applied the low-dimensional features to  $k$ -means clustering and reported the average results of 50 independent runs of the  $k$ -means clustering. Clustering accuracy was measured by normalized mutual information (NMI) (Strehl and Ghosh, 2003)

<sup>3</sup> The word “large-scale” means that the data cannot be handled by traditional kernel methods on a common personal computer.

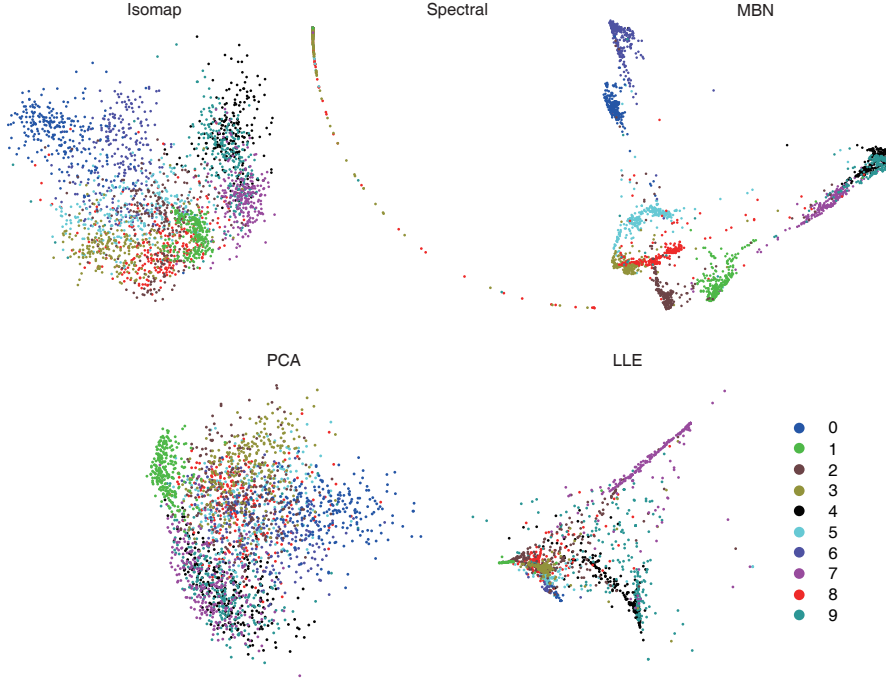


Fig. 6: Visualizations of a subset (5,000 images) of MNIST produced by 5 comparison methods and MBN at layer 9. For clarity, only 250 images per digit are drawn. Visualizations produced by other layers of MBN are shown in Fig. 18 in Appendix B.

which was proposed to overcome the label indexing problem between the ground-truth labels and the predicted labels. It is one of the standard evaluation metrics of clustering. It also has a strong one-to-one correspondence with classification accuracy in supervised learning.

#### 5.1.1 Experiment on small subsets of MNIST digits

We constructed 5 small subsets of MNIST. Each subset contains 5,000 randomly-sampled images with 500 images per digit. The recommended MBN setting in Section 2.1 for this data set was as follows. The number of hidden layers was set to 8. Parameters  $k$  from layer 1 to layer 8 were set to 2500-1250-625-312-156-78-39-19 respectively. Parameter  $V$  was set to 400 for all layers. Parameter  $a$  was set to 0.5. The similarity measurement at the bottom layer was Euclidean distance. The output dimension of EM-PCA was selected from  $\{2, 3, 5, 10, 20, 30\}$ .

We compared MBN with several dimensionality reduction methods, including PCA, isometric feature mapping (Isomap) (Tenenbaum et al, 2000), locally linear embedding (LLE) (Roweis and Saul, 2000), and spectral clustering (Spectral) (Ng et al, 2002) in data visualization and clustering. The parameters of Isomap and LLE (i.e., the number of neighboring data points) were searched from 2 to 20. The parameter of spectral clustering (i.e., the kernel width of the Gaussian RBF kernel) was searched through  $\{2^{-3}A, 2^{-2}A, \dots, 2^3A\}$  where  $A$  is the average pairwise Euclidean distance between data points.



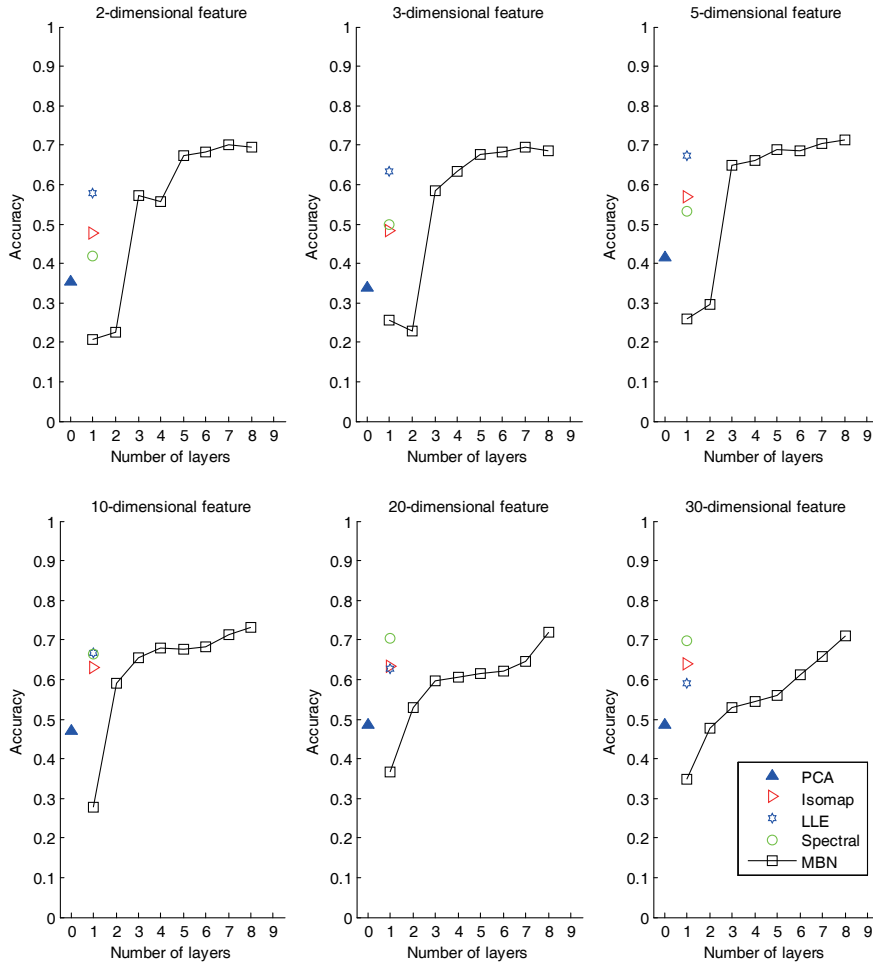


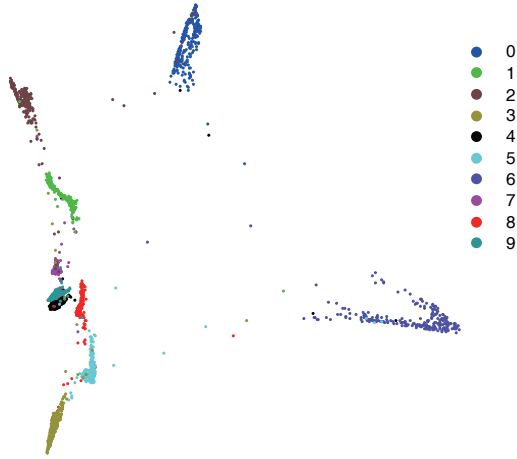
Fig. 7: Clustering accuracy (in terms of NMI) produced by dimensionality reduction methods on subsets (5,000 images per subset) of MNIST.

Table 2: Accuracy (in terms of NMI) of clustering algorithms on the 5,000 MNIST images.

$k$ -means clustering	GMM	Hierarchical clustering	MBN
48.9%	24.3%	63.0%	73.0%

The visualization result in Fig. 6 shows that the low-dimensional feature produced by MBN has the smallest within-class variance and largest between-class distance among the features produced by the comparison methods.

We summarized the average clustering accuracy on the subsets in Fig. 7. The clustering accuracy of MBN with 2, 3, 5, 10, 20, and 30 dimensional outputs is 69.4%, 68.6%, 71.4%, 73.0%, 71.8%, and 71.1% respectively, which demonstrates the robustness of MBN with respect to different output dimensions of EM-PCA. Note that the reported clustering accuracy



**Fig. 8** Visualization of the 10,000 test images of MNIST produced by MBN at layer 8. For clarity, only 250 images per digit are drawn. See Appendix B for the visualizations produced by other layers.

of MBN was produced from the top hidden layer; the performance at other hidden layers in Fig. 7 is merely for demonstrating the evolving process of the clustering accuracy.

We also compared with  $k$ -means clustering, GMM, and hierarchical clustering that ran with the 784-dimensional raw features. The number of natural classes was given to all methods. GMM was trained with a full covariance matrix. Agglomerative hierarchical clustering used the Ward’s minimum variance method in the Euclidean space, which was (to our knowledge) the best setting of hierarchical clustering on MNIST. The clustering accuracy in Table 2 shows that MBN outperforms the comparison methods. It is worthy noting that GMM performs poorly, since the distribution of the images is highly non-Gaussian and nonlinear.

### 5.1.2 Experiment on MNIST Digits

To investigate the scalability MBN, we ran experiments on the full MNIST handwritten digits. We trained models using the 60,000 training images, and evaluated their effectiveness on the 10,000 test images.

The recommended MBN setting in Section 2.1 for this data set was as follows. Parameters  $k$  from layer 1 to layer 10 were set to 8000-4000-2000-1000-500-250-125-62-31-15 respectively, where  $k_1 = 8000$  was the largest value our hardware can deal with at the time. Parameter  $V$  was set to 400 for all layers. Parameter  $a$  was set to 0.5. The similarity measurement at the bottom layer was Euclidean distance. The output dimension of EM-PCA was selected from  $\{2, 3, 5, 10, 20, 30\}$ . To investigate the effect of the network size of MBN on performance, we also ran a small-scale MBN with parameters  $k$  from layer 1 to layer 7 set to 1000-500-250-125-62-31-15.

The visualization result in Fig. 8 shows that MBN not only provides a clear visualization but also maintains the relative positions of the digits. For the task of clustering, we show the comparison result with PCA in Fig. 9. The test accuracy of MBN with 2, 3, 5, 10, 20, and 30 dimensional outputs is 74.9%, 79.6%, 81.4%, 82.8%, 82.9%, and 82.2% respectively. The test accuracy of the small-scale MBN with the 6 dimensional outputs is 69.3%, 72.3%, 73.8%, 76.4%, 77.6%, and 76.7% respectively. The result indicates that enlarging the network size of MBN improves its generalization ability.

We further compared MBN with  $k$ -means clustering, GMM, and hierarchical clustering that ran with the 784-dimensional raw features. We ran the comparison methods on the

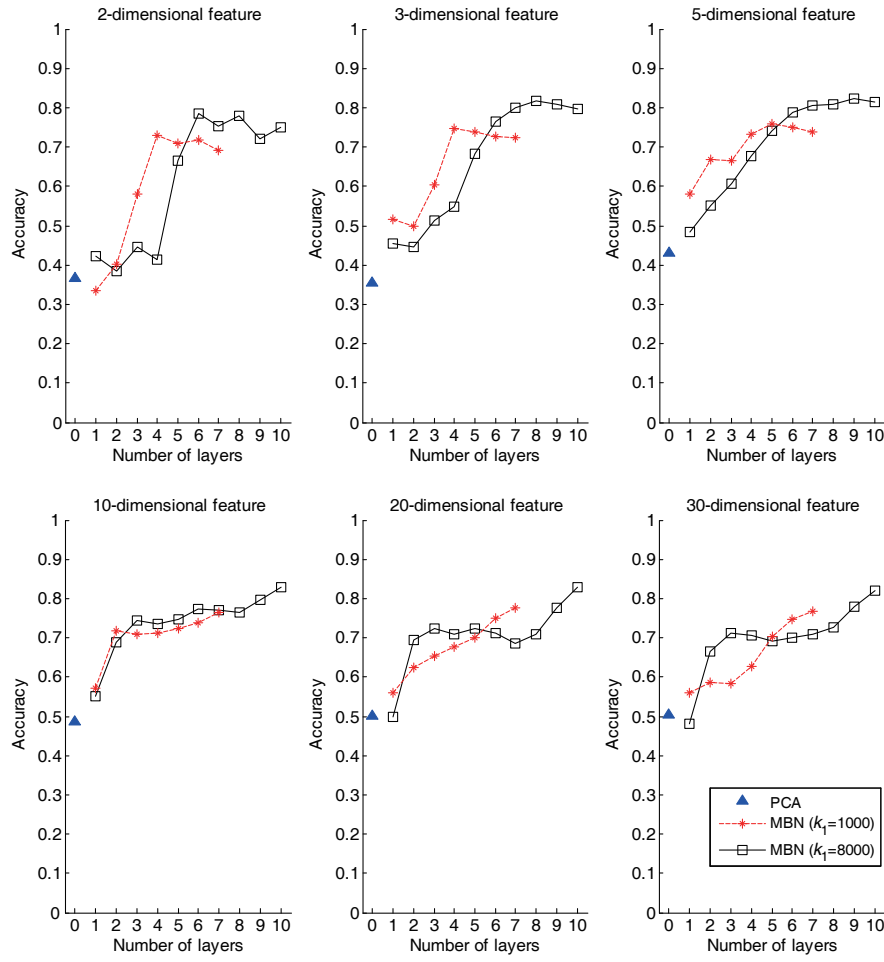


Fig. 9: Clustering accuracy (in terms of NMI) produced by MBN and PCA on the 10,000 test images of MNIST.

Table 3: Accuracy (in terms of NMI) of clustering algorithms on the 10,000 test images of MNIST.

$k$ -means clustering	GMM	Hierarchical clustering	MBN
50.6%	15.0%	71.2%	82.9%

10,000 test images directly and in the same experimental setting as in Section 5.1.1. The comparison result in Table 3 shows that MBN outperforms the comparison methods.

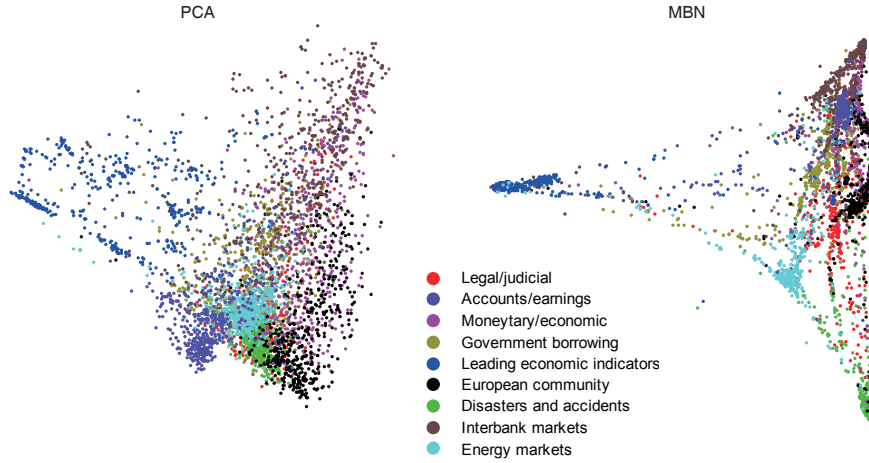


Fig. 10: Visualization of 9 demo topics of the Reuters newswire stories produced by LSA and the MBN at layer 8. For clarity, only 500 documents per topic are drawn.

## 5.2 Document retrieval

To further illustrate the scalability of MBN, we also compared MBN with latent semantic analysis (LSA) (Deerwester et al, 1990), a document retrieval method based on PCA, on a larger data set—Reuters newswire stories (Lewis et al, 2004) which consist of 804,414 documents. The data set of the Reuters newswire stories are divided into 103 topics. Because the topics are grouped into a tree structure, we only preserved the leaf topics. As a result, 82 topics remained, and there were 107,132 unlabeled documents. We preprocessed each document as a vector of 2,000 commonest word stems by the *rainbow* software (McCallum, 1998) where each entry of a vector was set to the word count and other parameters were set to their default values. We normalized each vector by dividing its entries by its  $\ell_2$ -norm.

For the task of visualizing the documents, we picked 9 demo topics, each with 5,000 documents. We adopted the recommended setting in Section 2.1. The number of hidden layers was set to 8. Parameters  $k$  from layer 1 to layer 8 were set to 2000-1000-500-250-125-62-31-15 respectively. Parameter  $V$  was set to 200. Parameter  $a$  was set to 0.5. The similarity measurement at the bottom layer was the same as other hidden layers. Fig. 10 shows that MBN produces a better visualization than LSA.

For the task of document retrieval, we randomly selected half of the data set for training and the other half for test. We recorded the average accuracy over all 402,207 queries in the test set at the document retrieval setting, where a query and its retrieved documents were different documents in the test set. If an unlabeled document was retrieved, it was considered as a mistake. If an unlabeled document was used as a query, no relevant documents would be retrieved, which means the precisions of the unlabeled query at all levels were zero. The similarity between two documents was measured by the Euclidean distance between the low-dimensional representations of the documents.

To investigate the effect of the network size of MBN on performance, we configured MBN with two parameter settings, both of which followed the recommended setting in Section 2.1. The first MBN set parameters  $k$  to 500-250-125 respectively; the second MBN set parameters  $k$  to 2000-1000-500-250-125 respectively. Other parameters of the two networks

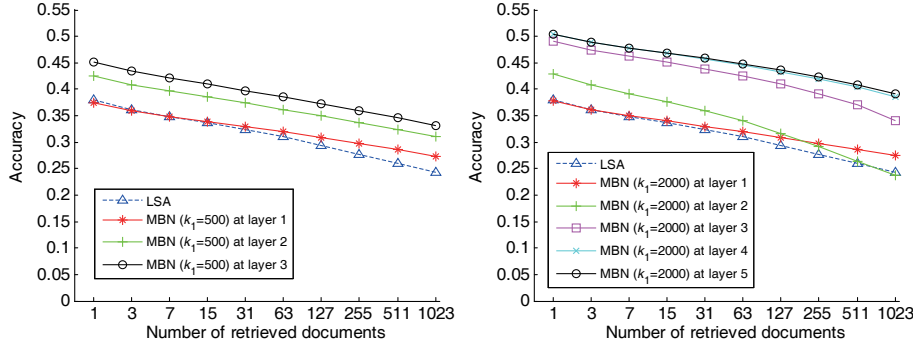


Fig. 11: Average accuracy curves of retrieved documents produced by MBN and LSA on the test set (82 topics) of the Reuters newswire stories.

Table 4: A basic hyperparameter setting for MNIST.

Parameter	Value
$L$	Automatically determined by $\{k_l\}_{l=1}^L$
$\{k_l\}_{l=1}^L$	$k_1 = 4500$ , $k_l = \delta k_{l-1}$ , and $k_L \geq 1.5c$ where $\delta = 0.5$ is the decay factor and $c = 10$ is the number of natural classes
$a$	0.5
$V$	400

were the same:  $V = 200$  and  $a = 0.5$ . The similarity measurement at the bottom layer was the same as other hidden layers. The output dimension of EM-PCA was 5.

Experimental results in Fig. 11 show that the MBN with the small network reaches an accuracy curve of over 8% higher than LSA; the MBN with the large network reaches an accuracy curve of over 13% higher than LSA. The results indicate that enlarging the network size of MBN improves its generalization ability.

### 5.3 Effects of hyperparameters on performance

In this subsection, we analyze the robustness of MBN to different hyperparameter settings, and demonstrate how the recommended parameter setting in Section 2.1 was selected.

To prevent an exhausting search over all possible combinations of the hyperparameters, we adopted a basic hyperparameter set in Table 4. When we investigated the effect of some hyperparameter on performance, we tuned the hyperparameter and kept other hyperparameters fixed as in Table 4.

We ran experiments on the MNIST handwritten digits. For each hyperparameter setting, we trained a model with the full training set (60,000 images) and evaluated it on the 10,000 test images. We took the output from the top layer of MBN as the input of EM-PCA, and used 5-dimensional features for clustering, unless otherwise stated. The clustering accuracy was evaluated by NMI.

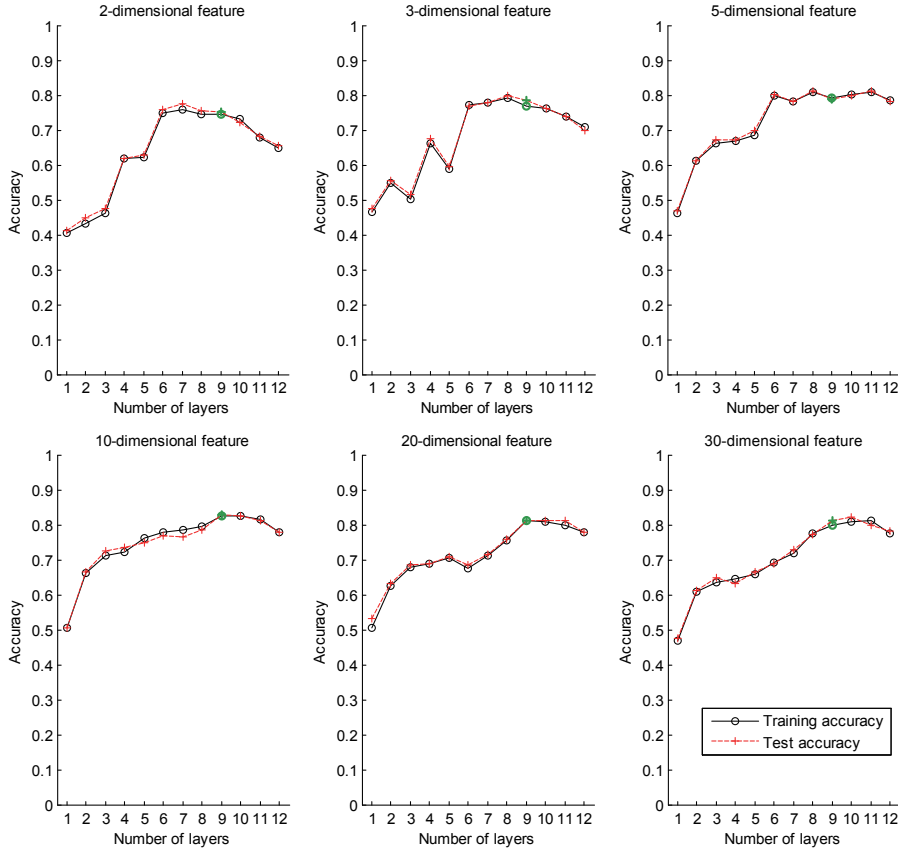


Fig. 12: Effect of parameter  $k_L$  on MNIST. The layer with  $k \approx 1.5c$  is marked in green color.

### 5.3.1 Effect of hyperparameter $k_L$

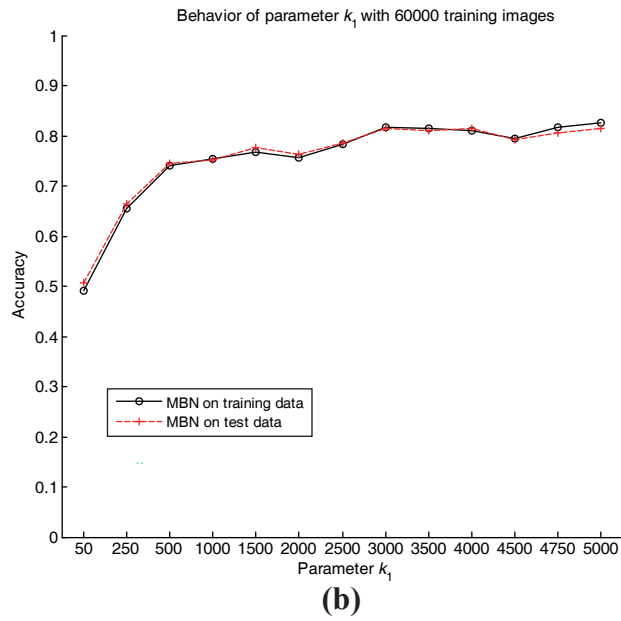
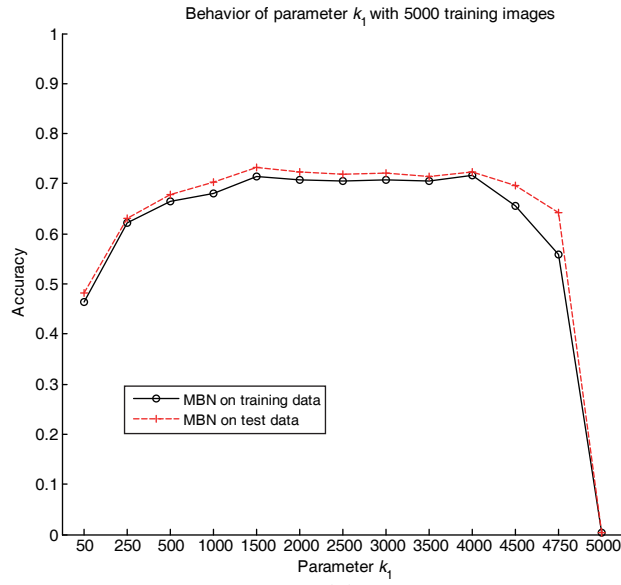
We modified Table 4 by setting  $k_L > 1$ . We selected the output dimension of EM-PCA from  $\{2, 3, 5, 10, 20, 30\}$ .

The experimental result with the full training set (Fig. 12) shows that the performance drops apparently when  $k < 1.5c$ . Therefore, we should constrain  $k_L$  to be at least larger than  $1.5c$ . The key idea behind  $k_L \geq 1.5c$  is that each  $k$ -centers clustering should be stronger than random guess (Schapire, 1990). It seems that the weakest meaningful learner should have  $k_L \geq c$ .

### 5.3.2 Effect of hyperparameters $k_1$

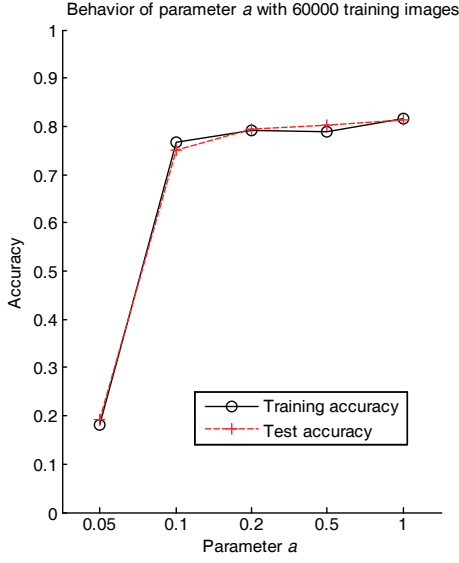
We modified Table 4 by selecting  $k_1$  from  $\{50, 250, 500 : 500 : 4500, 4750, 5000\}$ , where the symbol  $500 : 500 : 4500$  denotes a serial parameter values starting from 500 with an increment of 500 and ending up with 4500.

We first show the experimental result with a small-scale training set that consists of 5,000 images in Fig. 13a. From the figure, we observe the following phenomena. The accu-



**Fig. 13** Effect of parameter  $k_1$  on MNIST.

racy is improved gradually when  $k_1 \leq 1500$ , remains unchanged when  $1500 \leq k_1 \leq 4000$ , and drops sharply when  $k_1 \geq 4000$ . This phenomenon can be explained by the geometric principle of MBN in Section 3.1. Specifically, when  $k_1 \leq 1500$ , a group of base clusterings can build a local coordinate system for each input data point successfully. When  $k_1$  becomes larger, the local coordinate systems become more localized around the input training data points (i.e. more accurate). However, when  $1500 \leq k_1 \leq 4000$ , a training data point may



**Fig. 14** Effect of parameter  $a$  on MNIST

be selected by many base clusterings, which not only increases the correlation between the base clusterings but also decreases the effectiveness of the base clusterings. These negative factors offset the merit of enlarging  $k_1$ , which results in a stable accuracy curve. When  $k_1 \geq 4000$ , the negative factors become dominant, which results in a significant performance drop. For example, when  $k_1 = 5000$ , all base clusterings not only are identical but also provide no information for the data distribution, since the nearest centers of any input training data point are the data point itself. Because MBN performs robustly when  $k_1 \in [0.3n, 0.8n]$ , we recommend  $k_1 = \min\{0.5n, k_{\max}\}$ .

We show the experimental result with the full training set in Fig. 13b. From the figure, we observe that the accuracy is improved with increasing  $k_1$ , and the training and test accuracy curves match well.

### 5.3.3 Effect of hyperparameter $a$

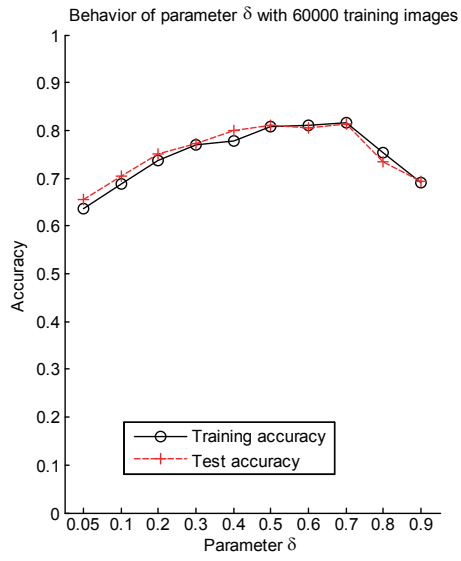
We modified Table 4 by selecting  $a$  from  $\{0.05, 0.1, 0.2, 0.5, 1\}$ . The experimental results show that the performance is generally robust when  $a \geq 0.1$ , and drops significantly when  $a = 0.05$ . Therefore, we should prevent setting  $a$  to a very small value. Empirically, we recommend setting  $a = 0.5$ , since it not only reduces the computational complexity but also is far from unsafe values.

### 5.3.4 Effect of decay factor $\delta$

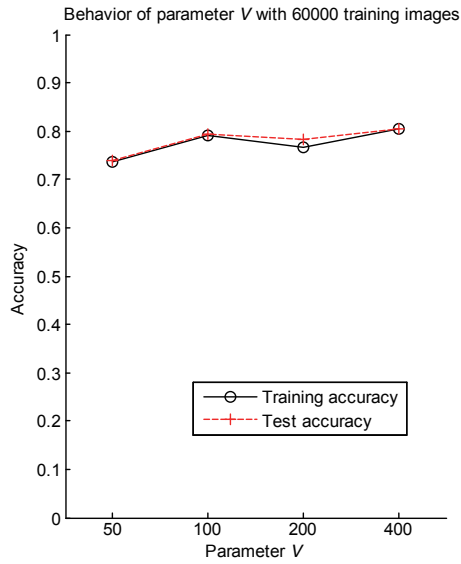
Given  $k_1$  and  $k_L$ , the decay factor  $\delta$  controls the convergence speed from  $k_1$  to  $k_L$ , which in turn determines the value of parameter  $L$ . We modified Table 4 by selecting  $\delta$  from  $\{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ .

The experimental result in Fig. 15 shows that the performance is sensitive to  $\rho$ ; the best performance happens at  $\rho \in [0.5, 0.7]$ . When  $\rho \leq 0.3$ , the performance drops significantly. The reason this happens is that the setting is too aggressive to satisfy the locally linear assumption. When  $\rho \geq 0.8$ , the performance also drops significantly. This setting is





**Fig. 15** Effect of decay factor  $\delta$  on MNIST.



**Fig. 16** Effect of parameter  $V$  on MNIST.

theoretically correct, however, the model is too subtle that it cuts the data points in the same class into many small pieces. Hence, we recommend setting  $\rho = 0.5$ .

### 5.3.5 Effect of hyperparameter $V$

We modified Table 4 by selecting  $V$  from  $\{50, 100, 200, 400\}$ . The experimental result in Fig. 16 shows that the performance is robust when  $V \geq 100$ . Therefore, we recommend setting  $V \geq 100$ .

Based on the above analysis, we summarized a typical hyperparameter setting in Section 2.1. We also found that the MBN with the recommended hyperparameter setting worked well on many real-world data sets, such as those in Appendix E.

## 6 Conclusions and future work

In this paper, we have proposed multilayer bootstrap network for nonlinear dimensionality reduction. MBN has a novel network structure that each expert is a  $k$ -centers clustering whose centers are randomly sampled data points with randomly selected features; the network is gradually narrowed from bottom up.

MBN is easily understood, implemented, and used, in respect of both the algorithm itself and the interpretation of its geometric principle. It learns an invariant representation of data by implicitly building a vast number of hierarchical trees whose number is exponentially larger than the number of the base  $k$ -centers clusterings. It adopts an adaptive similarity metric of data, which provides a novel way for the similarity metric problem of unsupervised learning. Its effectiveness can be proved from the perspective of the bias-variance decomposition theory.

MBN performs robustly with a wide range of parameter settings. Its time and storage complexities scale linearly with the size of training data. It supports parallel computing naturally. Empirical results demonstrate its efficiency at the training stage and its effectiveness in data visualization, clustering, and document retrieval. MBN extended data resampling and model ensembling methods to an unsupervised multilayer architecture.

In the appendix, we have further proposed compressive multilayer bootstrap network to compress the network size of MBN. It not only inherits the effectiveness of MBN on unsupervised learning but also inherits the effectiveness and efficiency of neural networks on supervised learning for its effectiveness and efficiency on unsupervised learning. It is a general framework of unsupervised model compression.

There is much work to do in the future. For example, (i) algorithms that further decorrelate the base clusterings by new randomization steps may be discovered. We have found in our preliminary study that applying a crossover-like random reconstruction operation in evolutionary computing, which randomly exchanges multiple features of the centers of a clustering that belongs to the same ground-truth class, can produce much more compact low-dimensional representations than the results reported in this paper. Although it is difficult to identify the centers that belong to the same ground-truth class, it provides a promising sign that more compact representations may be achievable by exploiting new randomization steps. (ii) More theoretical analyses are needed. Motivated by the analysis of the generalization ability of supervised learning, we have proved that MBN does not lose information as much as it appears. We have analyzed briefly the connection between MBN and the  $\ell_1$ -norm regularized sparse coding. We have also observed empirically the connection between the model complexity and the performance in Fig. 13a. Is it possible to generalize the well studied supervised learning theory to MBN, and more generally unsupervised learning? (iii) More applications are needed. Besides the application to unsupervised clustering and document retrieval, we have applied a single-layer bootstrap network to supervised speaker recognition, and observed some encouraging result comparing to a standard GMM-based method.

## Acknowledgements

The author thanks Prof DeLiang Wang for providing the Ohio Supercomputing Center, Columbus, OH, USA for the empirical study, and Dr Yuxuan Wang for providing his NN code.

## A Complexity analysis

**Theorem 3** *The computational and storage complexities of MBN at the bottom layer are:*

$$O_{time} = O(dskVn) \quad (19)$$

$$O_{storage} = O((ds+V)n+kV) \quad (20)$$

*and the computational and storage complexities of MBN at other layers are:*

$$O_{time} = O(kV^2n) \quad (21)$$

$$O_{storage} = O(2Vn+kV) \quad (22)$$

which scale linearly with respect to the size of the data set  $n$ , where  $d$  is the dimension of the original feature,  $s$  is the sparsity of the data (i.e., the ratio of the non-zero elements over all elements), and other hyperparameters of MBN are described in Table 1.

If MBN is not used for prediction, then MBN needs not to be saved, which further reduces the storage complexity to  $O((ds+V)n)$  at the bottom layer and  $O(2Vn)$  at other layers.

*Proof* For training the  $l$ th hidden layer, we suppose that the input is a  $d^{(l)}$ -dimensional data set that contains  $n$  data points; the sparsity of the input is  $s^{(l)}$ ; each layer contains  $V$  clusterings; and the number of the output units of each clustering at the  $l$ th layer is  $k^{(l)}$ . It is easy to derive the computational complexity of the hidden layer as  $O_{time} = O(d^{(l)}ns^{(l)}k^{(l)}V)$ . If  $l \neq 1$  (i.e., the layer is not the bottom one), we have  $d^{(l)} = Vk^{(l-1)}$ ,  $s^{(l)} = 1/k^{(l-1)}$ , and can further derive the computational complexity of the  $l$ th layer as  $O_{time} = O(nk^{(l)}V^2)$ .

We need an  $O((d^{(l)}s^{(l)} + d^{(l+1)}s^{(l+1)})n)$  space to store the input and output. We need at most an  $O(k^{(l)}V)$  space to store the model, since we only need to remember the indices of the centers in the input data and the indices of the dimensions of the centers that are randomly shifted. Summing the two items equals to  $O((d^{(l)}s^{(l)} + d^{(l+1)}s^{(l+1)})n + k^{(l)}V)$ . Substituting the relation  $d^{(l+1)} = Vk^{(l)}$  and  $s^{(l+1)} = 1/k^{(l)}$  to the summation reaches a conclusion:  $O_{storage} = O((d^{(l)}s^{(l)} + V)n + k^{(l)}V)$ . If  $l \neq 1$ , then  $O_{storage} = O(2Vn + k^{(l)}V)$ . If  $k^{(l-1)} = 2k^{(l)}$  which is a typical setting, then  $O_{storage} = O(2Vn + k^{(l)}V)$ . Theorem 3 is proved.  $\square$

Fortunately, the empirical time complexity does not grow with  $O(V^2)$  but with  $O(V)$ . Specifically, the running time with different  $V$  on MNIST are shown in Fig. 17a. From the figure, we find that the time complexity scales linearly but not squarely with  $V$ , which conflicts with our theoretical conclusion on the time complexity of MBN. Hence, we further draw the training time of each layer in Fig. 17b. From the figure, we find that training the bottom layer is the most time-consuming part which consumes 80% of the total running time when  $V = 50$  and 50% of the total running time when  $V = 400$ .

We further recorded the total running time of MBN and the running time of MBN on training each layer on the Reuters newswire stories in Figs. 17c and 17d respectively. From the figures, we find that the running time scales linearly with  $V$  too, even though the running time on training the bottom layer does not dominate the total running time.

Summarizing the above experimental phenomena, the only explanation for why the empirical time complexity scales linearly but not squarely with  $V$  is that the input data is sparse. Specifically, the multiplication of two sparse matrices only considers the element-wise multiplication of two elements that are both nonzero, as a result, when the input data is sparse, one factor  $V$  is offset by the sparsity factor  $s$ .

## B Supplementary figures

Fig. 18 is the supplement to Fig. 8 in the main text.

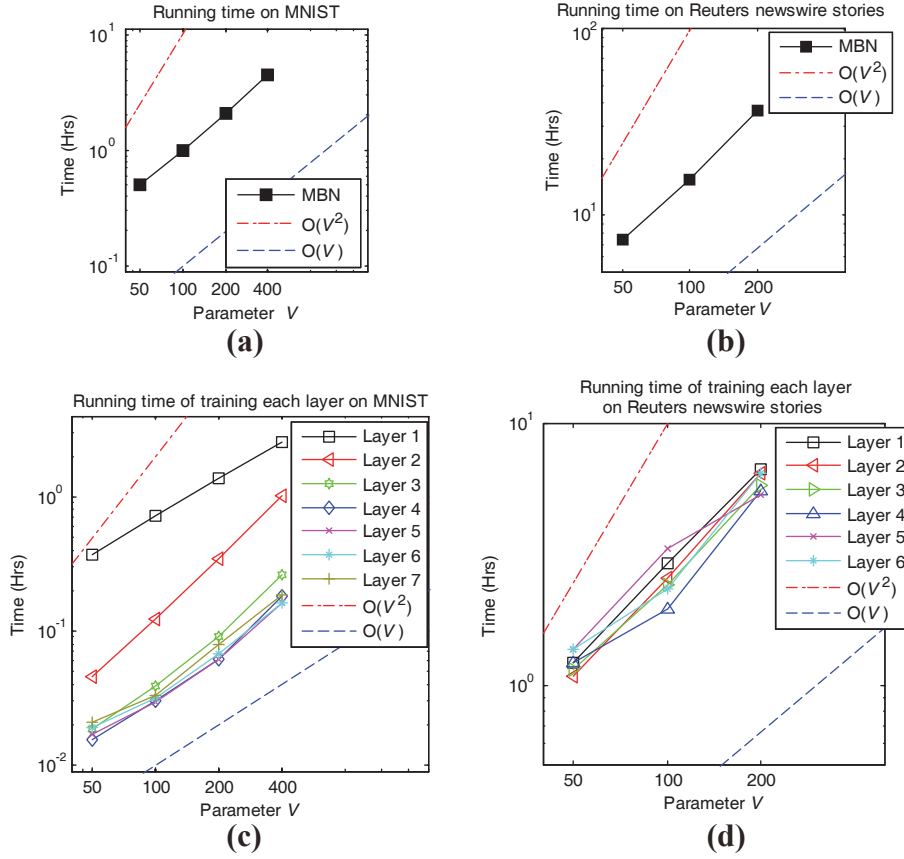


Fig. 17: Time complexity analysis on parameter  $V$ .

### C Source code and project page

Downloadable from <https://sites.google.com/site/zhangxiaolei321/mbn>

### D Supplementary algorithm: Compressive multilayer bootstrap networks

The network size of MBN is large, which is inefficient for prediction. To alleviate this problem, we propose *compressive multilayer bootstrap network* (compressive MBN) (Fig. 19) as follows:

- **MBN training.** The first step trains MBN with a give training set, and outputs a low dimensional representation of the training data.
- **Application [optional].** The second step applies the low dimensional representation to a given application in unsupervised learning, and outputs a prediction of the training data.
- **Neural network (NN) training.** The third step trains a NN with the original feature of the training data as the input and the prediction result as the target. Finally, the NN model is used for prediction.

After model compression, the prediction time complexity is reduced from  $O(kV^2L)$  per data point to  $O(zq)$  per data point where  $z \ll kV^2$  denotes the number of computational units of NN per layer and  $q < L$  denotes the number of layers of NN. Hence, compressive MBN not only inherits the effectiveness of MBN on unsupervised learning but also inherits the effectiveness and efficiency of neural networks on supervised

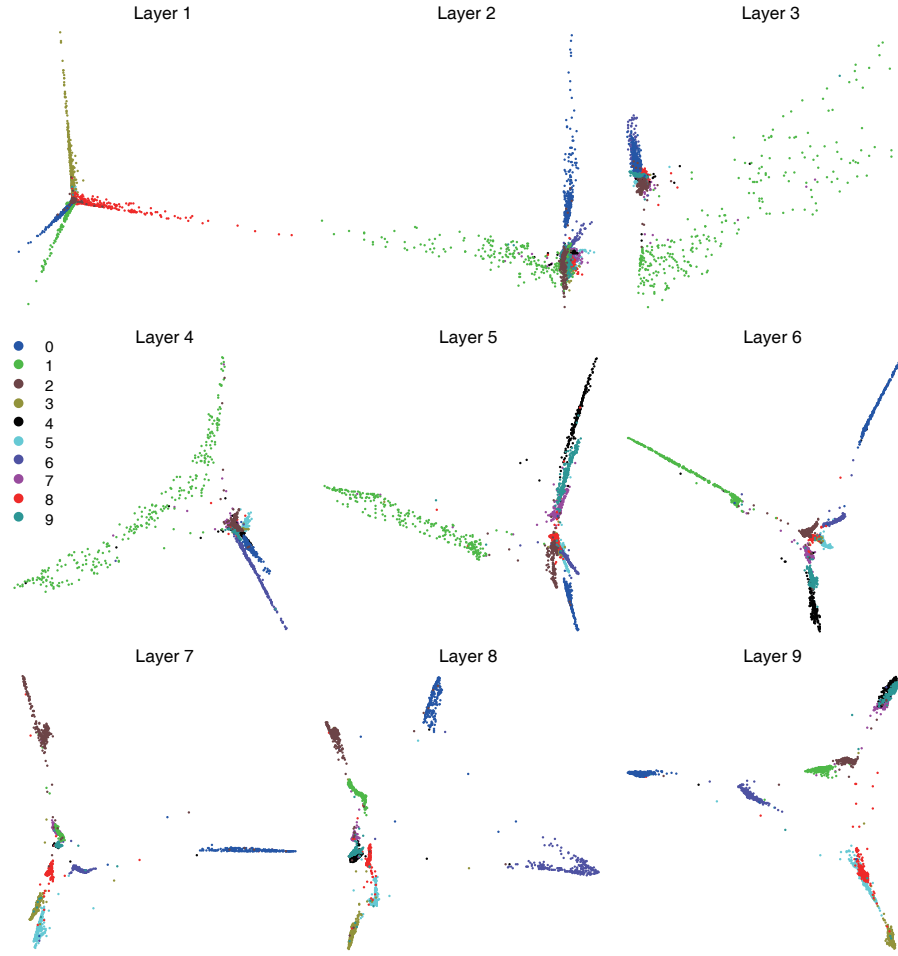
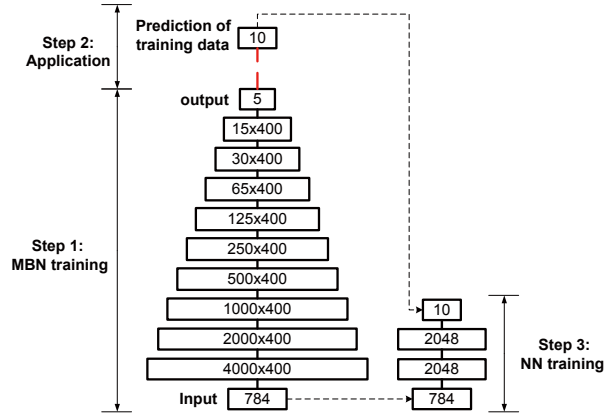


Fig. 18: Visualizations of MNIST produced by MBN at different layers.

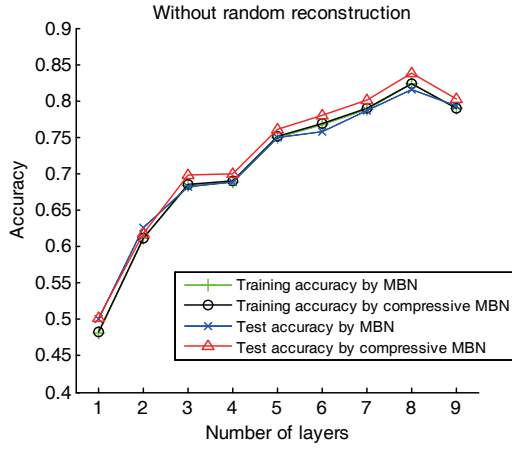
learning for its effectiveness and efficiency on unsupervised learning. Empirically, compressive MBN speeds up the prediction process by over thousands of times.

Compressive MBN is a general framework for unsupervised model compression, though it aims to alleviate the prediction complexity of MBN. We may modify the first step by using other unsupervised methods. We may use other supervised learners to compress the model size in the third step. We may also modify the second step for different applications. Some examples are as follows. (i) When compressive MBN is used for data visualization, we may omit the second step, and take the input and output of MBN as the input and output of NN respectively. (ii) When compressive MBN is used for unsupervised prediction, we may run a hard clustering algorithm on training data, and take the predicted indicator vectors as the pseudo “ground-truth” labels. For example, if a data point is assigned to the second cluster, then its predicted indicator vector is  $[0, 1, 0, 0, \dots, 0]$ .

From Appendix D.1 to D.3, we study compressive MBN in data visualization, unsupervised prediction, and document retrieval, comparing to MBN.



**Fig. 19** Principle of compressive MBN.



**Fig. 20** Comparison of the generalization ability of MBN and compressive MBN in clustering the MNIST images. The prediction time of MBN on the 10,000 test images is 4857.45 seconds. The prediction time of compressive MBN is 1.10 seconds.

### D.1 Generalization ability in image clustering

We studied the generalization ability of compressive MBN in clustering the 10,000 test images of MNIST (Lecun et al, 2004). We used all 60,000 training images of MNIST for unsupervised model training.

The parameter setting of MBN was as follows. Parameter  $V$  was set to 400. Parameters  $k$  from layer 1 to layer 9 were set to 4000-2000-1000-500-250-125-65-30-15 respectively. Parameter  $a$  was set to 0.5. The output dimension of EM-PCA was set to 5. We further encoded the 5-dimensional representations to 10-dimensional predicted indicator vectors by  $k$ -means clustering. As we can see, MBN is a large sparse model:  $784-(4000-2000-1000-500-250-125-65-30-15) \times 400-5$ .

For the compressive MBN training, we took the 784-dimensional original feature of the training images as the input of NN, and took their 10-dimensional predicted indicator vectors as the training target of NN. The parameter setting of NN was as follows. The network of NN was set to 784-2048-2048-10. The dropout rate was set to 0.2. The rectified linear unit was used as the hidden unit, and the sigmoid function was used as the output unit. The number of training epochs was set to 50. The batch size was set to 128. The learning rate was set to 0.001.

Because  $k$ -means clustering suffers from local minima, we ran the aforementioned methods 10 times and recorded the average results. The experimental results in Fig. 20 show that the training accuracy curves produced by MBN are completely consistent with those produced by compressive MBN; the test accuracy of compressive MBN is slightly higher than that of MBN. The most advanced property of compressive MBN is that the prediction efficiency is accelerated by around 4000 times.

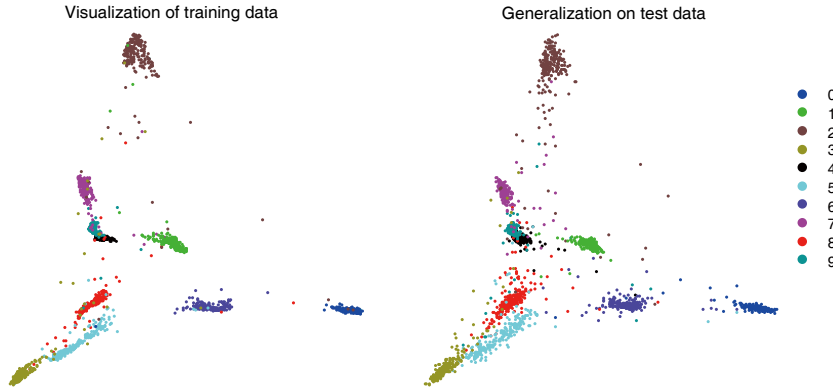


Fig. 21: Visualization of MNIST by the compressive MBN. The prediction time is similar with that in Section D.1.

Table 5: Accuracy comparison (in terms of NMI) between MBN and compressive MBN when using the 2-dimensional features of MNIST for clustering.

Methods	Training	Test
MBN	79.5%	79.6%
Compressive MBN	79.9%	79.7%

## D.2 Generalization ability in image visualization

We studied the generalization ability of compressive MBN in visualizing the MNIST images. For the MBN training, parameters  $k$  from layer 1 to layer 8 were set to 4000-2000-1000-500-250-125-65-30. Other parameters were the same as in Section D.1.

For the compressive MBN training, we used NN to learn a mapping function from the original feature to a 2-dimensional representation produced by MBN. The network of NN was set to 784-2048-2048-2. The rectified linear unit was used as the hidden unit, and the linear function was used as the output unit. The dropout function was disabled. The batch size was set to 32. The number of training epoches was set to 50. The learning rate was set to 0.001.

The visualization results in Fig. 21 show that compressive MBN produces an identical 2-dimensional feature as MBN on the training set and generalizes well on the test set. The clustering result in Table 5 shows that, when the 2-dimensional features were applied to  $k$ -means clustering, MBN and compressive MBN produce similar accuracy on both the training and test sets.

## D.3 Generalization ability in document retrieval

We studied the generalization ability of compressive MBN in retrieving the Reuters newswire stories (Lewis et al, 2004). See Section 5.2 for the introduction and preprocessing of the data set. The parameter setting of MBN was the same as the model with  $k_1 = 2000$  in Section 5.2.

For the compressive MBN training, we used NN to learn a mapping function from the 2000-dimensional original feature to the 5-dimensional representation produced by MBN. The parameter setting of NN was as follows. The network was set to 2000-2048-2048-5. The dropout rate was set to 0.2. The rectified linear unit was used as the hidden unit, and the linear function was used as the output unit. The number of training epoches was set to 50. The batch size was set to 512. The learning rate was set to 0.001.

The result in Fig. 22 shows that (i) compressive MBN produces an almost identical accuracy curve as MBN at the top layer; (ii) when the prediction results of MBN are squeezed at a small dense region (e.g., the

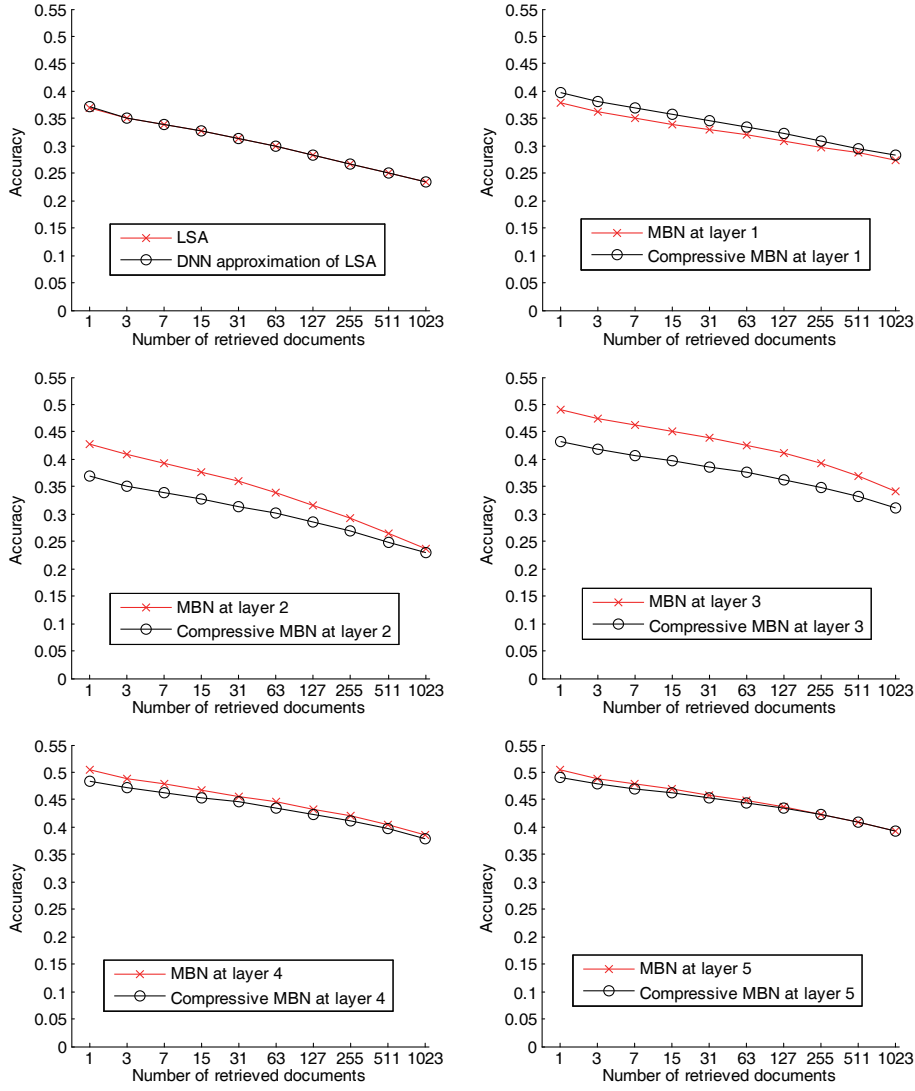


Fig. 22: Comparison of the generalization ability of MBN and compressive MBN on retrieving the Reuters newswire stories. The prediction time of MBN for producing the 5-dimensional feature of the 402,207 test documents is 190,574.74 seconds. The prediction time of compressive MBN for producing the feature is 88.80 seconds.

2-dimensional features at layers 2 and 3 in Fig. 18), compressive MBN may not produce an identical result with MBN (e.g., the accuracy curves at layers 2 and 3 in Fig. 22), however, this is usually not a big problem, since the prediction result produced by MBN at the top layer is usually not squeezed at a small and dense region; (iii) compressive MBN speeds up the prediction process by over 2000 times.



## E Supplementary results on clustering

Here we report results of an MBN-based clustering algorithm on 16 benchmark data sets as a supplement to the main text where we studied the property of MBN on the data sets of MNIST and RCV1. The details of the data sets are given in Table 6. The source code of the clustering algorithm, data sets, and more results are available at <https://sites.google.com/site/zhangxiaolei321/mbn>.

The 16 data sets cover topics in speech processing, chemistry, biomedicine, image processing, and text processing. “Extended-YaleB”, which contains 2,414 faces, is a post-processed version of the original extended Yale face database B. “20-Newsgroups”, which originally has 20,000 documents, was post-processed to a corpus of 18,846 documents, each belonging to a single topic. Because 45 topics of the original “Reuters-21578” contain few documents, we constructed a new corpus “Reuters-21578-Top20” that kept only the largest 20 topics. Similarly, because 66 topics of “TDT2” contain few documents, we constructed “TDT2-Top30” that included only the largest 30 topics.

Regarding the MBN-based clustering algorithm, **different** from the main text where each clustering result was an average of 50 independent runs of the  $k$ -means clustering applied to the low dimensional output of MBN, here we picked the clustering result that corresponded to the optimal objective value (i.e. the minimum mean square error) of the  $k$ -means clustering among the 50 candidate objective values as the final result of the MBN-based clustering. We used the default parameter setting in Section 2.1 without tuning (parameter  $k_1 = 8000$  for “MNIST(full)” and  $k_1 = 0.5n$  for all other data sets). The output dimension of MBN was set to the ground-truth number of classes. For the text corpora, we first preprocessed each document by dividing its entries by its  $\ell_2$ -norm, and then set the similarity metric of two documents at the bottom layer of MBN, say  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , as  $\mathbf{x}_1^T \mathbf{x}_2 / (\|\mathbf{x}_1\|_2 \|\mathbf{x}_2\|_2)$ . The entire process equals to the use of the cosine similarity metric  $\mathbf{x}_1^T \mathbf{x}_2 / (\|\mathbf{x}_1\|_2 \|\mathbf{x}_2\|_2)$ , which is the most common similarity metric in text processing. For other corpora, the similarity metric at the bottom layer was the Euclidean distance. We reported the average performance of 10 independent runs.

We compared with two common clustering methods,  $k$ -means clustering and PCA plus  $k$ -means clustering. For the data sets with IDs from 1 to 13, the PCA+ $k$ -means method preserved the top 98% largest eigenvalues and their corresponding eigenvectors. For the data sets with IDs from 14 to 16, PCA+ $k$ -means reduced the high-dimensional documents to 100-dimensional subspaces. For each data set, we ran  $k$ -means clustering with either the original feature or the low-dimensional output of PCA 50 times and recorded the result that corresponds to the optimal objective value among the 50 candidate values. We ran each comparison method 10 times and reported the average performance.

We evaluated the clustering result in terms of both NMI and clustering accuracy. The label indexing problem of clustering accuracy, i.e. the best permutation and matching between the predicted labels and the ground-truth labels, is found by the Hungarian method.

The clustering results in Tables 7 and 8 show that the MBN-based clustering outperforms the two comparison methods on 13 data sets, reaches tied performance with the  $k$ -means clustering on the Isolet1 data set, and is slightly worse than PCA+ $k$ -means on two text corpora where PCA+ $k$ -means uses 100 dimensional data to reach this performance.

At last, we note that although the default setting of MBN works well generally, the empirical performance of the MBN+ $k$ -means algorithm on many data sets can be further improved by controlling the network complexity modestly around the recommended setting. For example, if we set  $k_1 = 0.7n$  and  $\delta = 0.8$ , then NMI jumped to around 94% on COIL20 and around 85% on UMIST.

Table 6: Description of data sets.

ID	Name	# data points	# dimensions	# classes	Attribute
1	Isolet1	1560	617	26	Speech data
2	Wine	178	13	3	Chemical data
3	New-Thyroid	215	5	3	Biomedical data
4	Dermatology	366	34	6	Biomedical data
5	Lung-Cancer	203	12600	5	Biomedical data
6	COIL20(64x64)	1440	4096	20	Images
7	COIL100	7200	1024	100	Images
8	MNIST(small)	5000	768	10	Images (handwritten digits)
9	MNIST(full)	70000	768	10	Images (handwritten digits)
10	USPS	11000	1024	10	Images (handwritten digits)
11	UMIST	575	1024	20	Images (faces)
12	Extended-YaleB	2414	1024	38	Images (faces)
13	ORL	400	1024	40	Images (faces)
14	20-Newsgroups	18846	26214	20	Text corpus
15	Reuters-21578-Top20	7800	18933	20	Text corpus
16	TDT2-Top30	9394	36771	30	Text corpus

Table 7: NMI on 16 data sets. The number after  $\pm$  is the standard deviation.

		<i>k</i> -means	PCA+ <i>k</i> -means	MBN+ <i>k</i> -means
1	Isolet1	<b>77.21%</b> $\pm$ <b>0.92%</b>	56.74% $\pm$ 0.75%	<b>77.89%</b> $\pm$ <b>0.81%</b>
2	Wine	42.88% $\pm$ 0.00%	40.92% $\pm$ 0.00%	<b>55.49%</b> $\pm$ <b>4.07%</b>
3	New-Thyroid	49.46% $\pm$ 0.00%	49.46% $\pm$ 0.00%	<b>68.80%</b> $\pm$ <b>5.03%</b>
4	Dermatology	9.11% $\pm$ 0.11%	59.50% $\pm$ 0.10%	<b>82.40%</b> $\pm$ <b>2.24%</b>
5	Lung-Cancer	48.59% $\pm$ 1.07%	49.17% $\pm$ 0.96%	<b>50.64%</b> $\pm$ <b>3.46%</b>
6	COIL20(64x64)	78.03% $\pm$ 1.14%	79.00% $\pm$ 1.35%	<b>84.31%</b> $\pm$ <b>2.10%</b>
7	COIL100	76.98% $\pm$ 0.27%	69.64% $\pm$ 0.45%	<b>81.66%</b> $\pm$ <b>0.59%</b>
8	MNIST(small)	49.69% $\pm$ 0.14%	27.86% $\pm$ 0.08%	<b>77.12%</b> $\pm$ <b>0.35%</b>
9	MNIST(full)	50.32% $\pm$ 0.12%	28.05% $\pm$ 0.07%	<b>89.91%</b> $\pm$ <b>0.07%</b>
10	USPS	43.63% $\pm$ 1.78%	43.00% $\pm$ 0.05%	<b>80.52%</b> $\pm$ <b>0.81%</b>
11	UMIST	65.36% $\pm$ 1.21%	66.25% $\pm$ 1.10%	<b>74.48%</b> $\pm$ <b>1.91%</b>
12	Extended-YaleB	12.71% $\pm$ 0.63%	16.54% $\pm$ 0.56%	<b>43.36%</b> $\pm$ <b>0.44%</b>
13	ORL	75.55% $\pm$ 1.36%	75.81% $\pm$ 1.17%	<b>79.22%</b> $\pm$ <b>0.84%</b>
14	20-Newsgroups	Timeout	22.49% $\pm$ 1.41%	<b>41.61%</b> $\pm$ <b>0.51%</b>
15	Reuters-21578-Top20	Timeout	<b>45.86%</b> $\pm$ <b>0.90%</b>	44.20% $\pm$ 0.55%
16	TDT2-Top30	Timeout	<b>70.94%</b> $\pm$ <b>1.30%</b>	67.24% $\pm$ 0.98%

Table 8: Clustering accuracy on 16 data sets.

		<i>k</i> -means	PCA+ <i>k</i> -means	MBN+ <i>k</i> -means
1	Isolet1	<b>61.47%±1.93%</b>	38.62%±0.99%	<b>61.13%±2.52%</b>
2	Wine	70.22%±0.00%	78.09%±0.00%	<b>81.91%±2.61%</b>
3	New-Thyroid	86.05%±0.00%	86.05%±0.00%	<b>93.02%±1.60%</b>
4	Dermatology	26.17%±0.28%	61.67%±0.26%	<b>82.81%±7.67%</b>
5	Lung-Cancer	54.83%±2.29%	55.52%±1.89%	<b>60.84%±4.97%</b>
6	COIL20(64x64)	65.42%±2.49%	69.15%±2.38%	<b>74.51%±4.06%</b>
7	COIL100	49.75%±1.31%	43.42%±1.21%	<b>57.38%±1.85%</b>
8	MNIST(small)	52.64%±0.14%	34.49%±0.10%	<b>82.36%±0.46%</b>
9	MNIST(full)	53.48%±0.11%	35.14%±0.11%	<b>95.90%±0.04%</b>
10	USPS	43.70%±2.84%	47.63%±0.05%	<b>83.74%±0.97%</b>
11	UMIST	43.20%±1.66%	43.44%±1.92%	<b>56.96%±3.73%</b>
12	Extended-YaleB	9.61%±0.52%	10.59%±0.49%	<b>28.53%±0.86%</b>
13	ORL	54.37%±2.41%	54.55%±2.81%	<b>59.68%±1.58%</b>
14	20-Newsgroups	Timeout	22.61%±1.29%	<b>46.57%±1.10%</b>
15	Reuters-21578-Top20	Timeout	<b>29.35%±2.09%</b>	27.09%±2.25%
16	TDT2-Top30	Timeout	<b>52.48%±1.09%</b>	50.04%±2.21%

## References

- Boyd SP, Vandenberghe L (2004) *Convex Optimization*. Cambridge University Press.
- Breiman L (1996) Bagging predictors. *Machine Learning*, 24(2):123–140.
- Breiman L (2001) Random forests. *Machine Learning*, 45(1):5–32.
- Canu S, Grandvalet Y, Guigue V, Rakotomamonjy A (2005) SVM and kernel methods Matlab toolbox. <http://asi.insa-rouen.fr/enseignants/~arakoto/toolbox/index.html>
- Coates A, Ng AY (2011) The importance of encoding versus training with sparse coding and vector quantization. In: *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, (pp 921–928).
- Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- Dietterich TG (2000) Ensemble methods in machine learning. In: *Multiple Classifier Systems*, Springer, Cagliari, Italy, (pp 1–15).
- Dudoit S, Fridlyand J (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099.
- Efron B (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7(1):1–26.
- Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap*. CRC press.
- Fern XZ, Brodley CE (2003) Random projection for high dimensional data clustering: A cluster ensemble approach. In: *Proceedings of the 20th International Conference on Machine Learning*, vol 3, (pp 186–193).
- Fred AL, Jain AK (2005) Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850.
- Freund Y, Schapire RE (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In: *Proceedings of the 2nd European Conference on Computational Learning Theory*, Barcelona, Spain, (pp 23–37).
- Friedman J, Hastie T, Tibshirani R, et al (2000) Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, 28(2):337–407.
- Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning*. Springer.
- Lecun Y, Cortes C, Burges CJC (2004) THE MNIST DATABASE of handwritten digits. <http://yann.lecun.com/exdb/mnist/index.html>
- Lewis DD, Yang Y, Rose TG, Li F (2004) RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.
- McCallum A (1998) Rainbow. <http://www.cs.cmu.edu/~mccallum/bow/rainbow>
- Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: Analysis and an algorithm. In: *Advances in Neural Information Processing Systems 14*, Vancouver, British Columbia, Canada, (pp 849–856).
- Roweis ST (1998) EM algorithms for PCA and SPCA. In: *Advances in Neural Information Processing Systems 10*, Denver, CO, (pp 626–632).
- Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.
- Schapire RE (1990) The strength of weak learnability. *Machine Learning*, 5(2):197–227.
- Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.
- Strehl A, Ghosh J (2003) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.
- Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.
- Vega-Pons S, Ruiz-Shulcloper J (2011) A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372.